THE UNIVERSITY OF QUEENSLAND

AUSTRALIA

Structural Variation in Cancer Genomes & the Identification of Personalized DNA-based Cancer Biomarkers

Kelly Quek

(BSc. (Hons.), The University of Queensland)

A thesis submitted for the degree of Doctor of Philosophy at

The University of Queensland in 2015

Institute for Molecular Bioscience

Abstract

Cancer is the result of the accumulation of genetic and epigenetic alterations in key genes, which ultimately lead to uncontrolled growth of mutated cells. Genetic alterations range from small base substitutions to large chromosomal structural rearrangements. Many cancer genomes carry tens to hundreds of somatic structural rearrangements, which may have functional consequences. However, the patterns and characterization of somatic rearrangements in human cancers are still in early stages. In this thesis, I describe the genomic landscape of somatic structural rearrangements in human pancreatic cancers. I then investigate the potential mechanisms involved in somatic rearrangement formation and test whether such rearrangements can be used as biomarkers to monitor patient therapy.

Chapter 1 - Introduction: This chapter is a broad overview of somatic structural rearrangements in human cancers, motivation for studying them, and their potential use for clinical applications.

Chapter 2 – A workflow to increase verification rate of somatic chromosomal rearrangements using high throughput next generation sequencing: I established a high throughput workflow to rapidly verify somatic structural variants and identify the exact location of breakpoints using benchtop next generation sequencing and computational tools. The workflow examined more than 300 predicted breakpoints and identified breakpoint location in more than 80% of somatic events at base level. The results demonstrated that next generation sequencing is comparable to the conventional Sanger sequencing and can complete the verification workflow in a shorter timeframe, enabling rapid validation of events.

Chapter 3 – Patterns of somatic breakpoints may indicate repair mechanisms that were active or absent during the generation of genomic rearrangements: This chapter describes the spectrum of somatic rearrangements and breakpoints detected in 120 primary pancreatic ductal adenocarcinomas genomes. The analysis includes characterization of the breakpoint junctions to infer which potential DNA repair processes are occurring. The results revealed that the majority of tumours exhibited repair of chromosome structural breaks using microhomology suggesting that NHEJ is the main

mechanism of DNA repair in pancreatic cancer. Tumours with *BRCA*1 *or BRCA*2 gene mutations or with a high contribution of a BRCA-like mutational signature showed a higher frequency of somatic breakpoints with microhomology length 1 to 5 bp and lower frequency of breakpoints with a blunt end (0 bp) when compared to tumours harbouring *BRCA* wild type or low BRCA mutational signature. The similarity in breakpoint characteristics between tumours with *BRCA* mutation and BRCA mutational signature reinforce previous findings that a subtype of pancreatic tumour might have deficiency in the HR pathway and could respond to PARP1 inhibitors. Furthermore, the analysis of the DNA sequence surrounding the breakpoints revealed strong signal of A+Ts rich regions suggesting that the formation of somatic rearrangements could also be mediated by either retrotransposition activity or chromosomal fragile sites. Taken together, the analyses of breakpoint junctions highlighted that the formation of somatic rearrangements in pancreatic carcinogenesis is complex – potentially with more than one mechanism active within a cancer genome.

Chapter 4 – Identification of personalized DNA-based biomarkers for pancreatic cancer and the assessment of whether they can be used to monitor tumour burden and response to chemotherapy: In this chapter, I first evaluated the performance of sequencing and PCR based methods to detect ctDNA. Subsequently, I quantified ctDNA in the serum or plasma of the three pancreatic cancer patients. Little or no detection of ctDNA was observed in the analysed serum samples. For one patient, a tumour specific rearrangement was detected in the serum, and this patient had already presented with metastatic disease. Based on the results, it was hypothesized that ctDNA released from pancreatic cancer might be limited by: i) nature of pancreatic cancer; ii) the physiological location of the pancreas which could influence the amount of ctDNA released in the circulation, as the fragmented tumour DNA might be cleared by the liver and therefore reduced the opportunity to detect ctDNA in the circulation; iii) volume of plasma or serum used to isolate cfDNA; iv) the status of disease progression. In this analysis, the ctDNA was detected only in the sample collected at late stage pancreatic cancer (after the diagnosed of liver metastasis). Thus, in the context of pancreatic cancer disease, the quantification of ctDNA might be more suitable for recurrent disease, which has metastasized from the pancreas.

Chapter 5: I concluded my findings throughout my studies in Chapter 2, 3 and 4 with a summary and future directions.

Declaration by author

This thesis is composed of my original work, and contains no material previously published or written by another person except where due reference has been made in the text. I have clearly stated the contribution by others to jointly-authored works that I have included in my thesis.

I have clearly stated the contribution of others to my thesis as a whole, including statistical assistance, survey design, data analysis, significant technical procedures, professional editorial advice, and any other original research work used or reported in my thesis. The content of my thesis is the result of work I have carried out since the commencement of my research higher degree candidature and does not include a substantial part of work that has been submitted to qualify for the award of any other degree or diploma in any university or other tertiary institution. I have clearly stated which parts of my thesis, if any, have been submitted to qualify for another award.

I acknowledge that an electronic copy of my thesis must be lodged with the University Library and, subject to the policy and procedures of The University of Queensland, the thesis be made available for research and study in accordance with the Copyright Act 1968 unless a period of embargo has been approved by the Dean of the Graduate School.

I acknowledge that copyright of all material contained in my thesis resides with the copyright holder(s) of that material. Where appropriate I have obtained copyright permission from the copyright holder to reproduce material in this thesis.

Publications during candidature

Peer-reviewed papers

Kelly Quek, Katia Nones, Ann-Marie Patch, Lynn Fink, Felicity Newell, Nicole Cloonan, David Miller, Muhammad Fudlullah, Karin Kassahn, Angelika Christ, Tim Bruxner, Senel Idrisoglu, Craig Nourse, Ehsan Nourbakhsh, Shivangi Wani, Anita Steptoe, Matthew Anderson, Oliver Holmes, Conrad Leonard, Darrin Taylor, Scott Wood, Qinying Xu, Australian Pancreatic Cancer Genome Initiative, Peter Wilson, Andrew V. Biankin, John V. Pearson, Nic Waddell, M. Grimmond. 2014. A workflow to increase verification rate of chromosomal structural rearrangements using high throughput next generation sequencing. **BioTechniques** *57*:3-38

Ann-Marie Patch*, Felicity Newel*, **Kelly Quek***, Katia Nones, Nicole Cloonan, Lynne Fink, Karin Kassahn, Muhammad Fudullah, David Miller, Suzanne Manning, Matthew Anderson, Oliver Holmes, Conrad Leonard, Darrin Taylor, Scott Wood, Qinying Xu, Australian Pancreatic Cancer Genome Initiative, Peter Wilson, Andrew V. Biankin, Nic Waddell, Sean M. Grimmond, John V. Pearson. qSV tool : An integrative, multi methodological approach for detection of somatic structural variants from paired end whole genome sequencing data. (In preparation)

Kelly Quek*, Ann-Marie Patch*, Katia Nones, Sean M. Grimmond, Nicola Waddell. Circulating cell free DNA as tumour biomarker. Review for International Journal of Molecular Medicine (In preparation)

Peter Bailey, David K. Chang, Katia Nones, Amber L. Johns, Ann-Marie Patch, Marie-Claude Gingras, David K. Miller, Angelika N. Christ, Tim J. C. Bruxner, Michael C. Quinn, Craig Nourse, L. Charles Murtaugh, Ivon Harliwong, Senel Idrisoglu, Suzanne Manning, Ehsan Nourbakhsh, Shivangi Wani, Lynn Fink, Oliver Holmes, Matthew J. Anderson, Stephen Kazakoff, Conrad Leonard, Felicity Newell, Nick Waddell, Scott Wood, Qinying Xu, Peter J. Wilson, Nicole Cloonan, Karin S. Kassahn, Darrin Taylor, **Kelly Quek**, Alan Robertson, Jianmin Wu, Mark Pinese, Mark J. Cowley, Marc D. Jones, Emily K. Colvin, Adnan M. Nagrial, Emily S. Humphrey, Lorraine A. Chantrill, Amanda Mawson, Jeremy Humphris, Angela Chou, Marina Pajic, Christopher J. Scarlett,, Andreia V. Pinho, Marc Giry-Laterriere, Ilse Rooman, Jaswinder S. Samra, James G. Kench, Jessica A. Lovell, Neil D. Merrett, Christopher W. Toon, Krishna Epari, Nam Q. Nguyen, Andrew Barbour, Nikolajs Zeps, Kim Moran-Jones, Nigel B. Jamieson, Janet S. Graham, Fraser Duthie,

Karin Oien, Jane Hair, Robert Grützmann, Anirban Maitra, Christine A. Iacobuzio-Donahue, Christopher L. Wolfgang, Richard A. Morgan, Rita T. Lawlor,, Vincenzo Corbo, Claudio Bassi, Borislav Rusev, Paola Capelli, Roberto Salvia Giampaolo Tortora, Debabrata Mukhopadhyay, Gloria M. Petersen, Australian Pancreatic Cancer Genome Initiative, Donna M. Munzy, William E. Fisher, James R. Eshleman, Ralph H. Hruban, Christian Pilarsky, Jennifer P. Morton, Owen J Sansom, Aldo Scarpa, Elizabeth A. Musgrove, David A Wheeler, Anthony J. Gill, Richard A. Gibbs, John V. Pearson, Nicola Waddell, Andrew V. Biankin, and Sean M. Grimmond. Molecular Pathology of Pancreatic Cancer. (**Nature**, under revision)

Ann-Marie Patch*, Elizabeth L. Christie*, Dariush Etemadmoghadam*, Dale W. Garsed*, Joshy George, Sian Fereday, Katia Nones, Prue Cowin, Kathryn Alsop, Peter J Bailey, Karin S Kasshan, Felicity Newell, Michael C. J. Quinn, Stephen Kazakoff, **Kelly Quek**, Charlotte Wilhelm-Benartzi, Ed Curry, Huei San Leong, The Australian Ovarian Cancer Study Group, Anne Hamilton, Linda Mileshkin, George Au-Yeung, Catherine Kennedy, Jillian Hung, Yoke-Eng Chiew, Paul Harnett, Michael Friedlander, Michael Quinn, Jan Pyman, Stephen Cordner, Patricia O'Brien, Jodie Leditschke, Greg Young, Kate Strachan, Paul Waring, Walid Azar, Chris Mitchell, Nadia Traficante, Heather Thorne, Mark Shackleton, David K Miller, Gisela Mir Arnau, Richard Tothill, Tim Holloway, Timothy Semple, Ivon Harliwong, Craig Nourse, Ehsan Nourbakhsh, Suzanne Manning, Senel Idrisoglu, Timothy J. C. Bruxner, Angelika N. Christ, Barsha Poudel, Oliver Holmes, Matthew Anderson, Conrad Leonard, Andrew Lonie, Nathan Hall, Scott Wood, Darrin F. Taylor, Qinying Xu, J. Lynn Fink, Nick Waddell, Ronny Drapkin, Euan Stronach, Hani Gabra, Bob Brown, Andrea Jewell, Shivashankar Hiriyur Nagaraj, Emma Markham, Peter J. Wilson, Jason Ellul, Orla McNally, Maria Doyle, Ravikiran Vedururu, Collin Stewart, Ernst Lengyel, John V. Pearson, Nicola Waddell, Anna deFazio**, Sean M. Grimmond** and David D.L. Bowtell**. Whole genome characterisation of chemo-resistant ovarian cancer. **Nature** 2015, 521:489-494

Nicola Waddell, Marina Pajic, Ann-Marie Patch, David K. Chang, Karin S. Kassahn, Peter Bailey, Amber L. Johns, David Miller, Katia Nones, **Kelly Quek**, Michael Quinn, Alan Robertson, Muhammad Fudlullah, Tim Bruxner, Angelika Christ, Ivon Harliwong, Senel Idrisoglu, Suzanne Manning, Craig Nourse, Ehsan Nourbakhsh, Shivangi Wani, Peter J. Wilson, Emma Markham, Nicole Cloonan, Matthew Anderson, Lynn Fink, Oliver Holmes, Stephen Kazakoff, Conrad Leonard, Felicity Newell, Sarah Song, Darrin Taylor, Nick

Waddell, Scott Wood, Christina Xu, Jianmin Wu, Mark Pinese, Mark J. Cowley, Marc D. Jones, Emily K. Colvin, Adnan M. Nagrial, Emily S. Humphrey, Lorraine A. Chantrill, Amanda Mawson, Jeremy Humphris, Angela Chou,Christopher J. Scarlett, Andreia V. Pinho, Ilse Rooman, Jaswinder S. Samra, James G. Kench, Jessica A. Lovell, Neil D. Merrett, Christopher Toon, Krishna Epari, Nam Q. Nguyen, Andrew Barbour, Nikolajs Zeps, Nigel B. Jamieson, Janet S. Graham, Robert Grützmann, Ralph H. Hruban, Anirban Maitra, Christine A. Iacobuzio-Donahue, Christopher L. Wolfgang, Richard A. Morgan, Rita T. Lawlor, Vincenzo Corbo, Claudio Bassi, Massimo Falconi, Giuseppe Zamboni, Giampaolo Tortora, Margaret A. Tempero, Australian Pancreatic Cancer Genome Initiative, Anthony J. Gill, James R. Eshleman, Christian Pilarsky, Aldo Scarpa, Elizabeth A. Musgrove, Robert L. Sutherland, John V. Pearson, Andrew V. Biankin and Sean M. Grimmond. Whole Genomes Redefine the Mutational Landscape of Pancreatic Cancer. **Nature** 2015, 518: 495-501.

Katia Nones, Nicola Waddell, Nicci Wayte, Ann-Marie Patch, Peter Bailey, Felicity Newell, Oliver Holmes, J. Lynn Fink, Michael Quinn, Yue Tang, Guy Lampe, **Kelly Quek**, Kelly Loffler, Suzanne Manning, Senel Idrisoglu, David Miller, Qinying Xu, Nick Waddell, Peter Wilson, Timothy Bruxner, Angelika Christ, Ivon Harliwong, Craig Nourse, Ehsan Nourbakhsh, Matthew Anderson, Stephen Kazakoff, Conrad Leonard, Scott Wood, Peter Simpson, Lynne Reid, Lutz Krause, Damian Hussey, David Watson, Reginald Lord , Derek Nancarrow, David Gotley, B. Mark Smithers, David Whiteman, Nicholas Hayward, Peter Campbell, John Pearson, Andrew Barbour, Sean M. Grimmond. 2014. Genomic catastrophes frequently arise in esophageal adenocarcinoma and drive tumorigenesis. **Nature Communication** *5*, 5224.

Kuljeet Singh Sandhue, Guoliang Li, Huay Mei Poh, **Yu Ling Kelly Quek**, Yee Yen Sia, Su Qin Peh, Fabianus Hendriyan Mulawadi, Joanne Lim, Mile Sikic, Francesca Menghi, Anbupalam Thalamuthu, Wing Kin Sung, Xiaoan Ruan, Melissa Jane Fullwood, Edison Liu, Peter Csermely, Yijun Ruan. 2012. Large-Scale Functional Organization of Long-Range Chromatin Interaction Networks. **Cell Reports**. 2:1207-1219

Conference abstracts

Kelly Quek, Katia Nones, Ann-Marie Patch, Lynn Fink, Felicity Newell, David Miller, Muhammad Fudlullah, Karin Kassahn, Angelika Christ, Tim Bruxner , Suzanne Manning, Ivon Harliwong, Senel Idrisoglu, Craig Nourse, Ehsan Nourbakhsh, Shivangi Wani,

Matthew Anderson, Oliver Holmes, Conrad Leonard, Darrin Taylor, Scott Wood, Qinying Xu, Australian Pancreatic Cancer Genome Initiative, Peter Wilson, Andrew V. Biankin, John V. Pearson, Nic Waddell, Sean M. Grimmond. Structural Variation in Pancreatic Cancer Genomes. Australasian Genomic Technologies Association (AGTA), Melbourne, Australia. 2014. Poster presentation. **Awarded for best student poster, AUD $250.**

Kelly Quek, Katia Nones, Ann-Marie Patch, Lynn Fink, Felicity Newell, David Miller, Muhammad Fudlullah, Karin Kassahn, Angelika Christ, Tim Bruxner, Suzanne Manning, Ivon Harliwong, Senel Idrisoglu, Craig Nourse, Ehsan Nourbakhsh, Shivangi Wani, Matthew Anderson, Oliver Holmes, Conrad Leonard, Darrin Taylor, Scott Wood, Qinying Xu, Australian Pancreatic Cancer Genome Initiative, Peter Wilson, Andrew V. Biankin, John V. Pearson, Nic Waddell, Sean M. Grimmond. Structural Variation in Cancer Genomes. 15th EMBL PhD Symposium, EMBL, Heidelberg, Germany. 2013. Poster presentation. **Awarded for best poster prize, €100.**

Kelly Quek, Katia Nones, Ann-Marie Patch, Nicole Cloonan, Felicity Newell, Lynn Fink, Angelika Christ, David Miller, Tim Bruxner, Australian Pancreatic Cancer Genome Initiative, Nic Waddell, Andrew V. Biankin, John V. Pearson, Sean M. Grimmond. A Workflow To Increase Verification Rate of Chromosomal Structural Rearrangements Using High Throughput Next Generation Sequencing. Australasian Genomic Technologies Association (AMATA), Surfers Paradise, Australia. 2013. Poster presentation.

Kelly Quek, Katia Nones, Ann-Marie Patch, Nicole Cloonan, Felicity Newell, Lynn Fink, Angelika Christ, David Miller, Tim Bruxner, Australian Pancreatic Cancer Genome Initiative, Nic Waddell, Andrew V. Biankin, John V. Pearson, Sean M. Grimmond. A Workflow To Increase Verification Rate of Chromosomal Structural Rearrangements Using High Throughput Next Generation Sequencing. Lorne Genome Conference, Mantra Lorne, Victoria, Australia. 2013. Poster presentation.

Publications included in this thesis

Publication incorporated as Chapter 2

Kelly Quek, Katia Nones, Ann-Marie Patch, Lynn Fink, Felicity Newell, Nicole Cloonan, David Miller, Muhammad Fudlullah, Karin Kassahn, Angelika Christ, Tim Bruxner, Senel Idrisoglu, Craig Nourse, Ehsan Nourbakhsh, Shivangi Wani, Anita Steptoe, Matthew Anderson, Oliver Holmes, Conrad Leonard, Darrin Taylor, Scott Wood, Qinying Xu, Australian Pancreatic Cancer Genome Initiative, Peter Wilson, Andrew V. Biankin, John V. Pearson, Nic Waddell, Sean M. Grimmond. 2014. A workflow to increase verification rate of chromosomal structural rearrangements using high throughput next generation sequencing. **BioTechniques** *57*:3-38

Contributor	Statement of contribution
Kelly Quek (Candidate)	Analysed the data (100%) Wrote the paper (70%) Performed laboratory experiments (50%)
Katia Nones	Wrote and edited paper (20%)
Nic Waddell	Wrote and edited paper (10%)
Ann-Marie Patch, Lynn Fink, Felicity Newell, Nicole Cloonan, Karin Kassahn	Designed the algorithms of the qAmplicon software (100%)
QCMG sequencing team: David Miller, Muhammad Fudlullah, Angelika Christ, Tim Bruxner, Senel Idrisoglu, Craig Nourse, Ehsan Nourbakhsh, Shivangi Wani, Anita Steptoe	Performed the laboratory experiments and next generation sequencing (50%)
QCMG informatics team: Matthew Anderson, Oliver Holmes, Conrad Leonard, Darrin Taylor, Scott Wood, Qinying Xu, John V. Pearson	Performed variant calling and data management of the next generation sequence data (100%)
Australian Pancreatic Cancer Genome Initiative, Andrew V. Biankin	Contributed tumour samples (100%)
Katia Nones, Peter Wilson, Nic Waddell Sean M Grimmond	Oversaw the work (100%)

Contributions by others to the thesis

Samples were collected and processed as part of the Australian Pancreatic Genome Initiative (http://www.pancreaticcancer.net.au/) and International Cancer Genome Consortium (ICGC). The Australian ICGC project was led by Prof. Sean Grimmond (Queensland Centre for Medical Genomics) and Prof. Andrew Biankin (Garvan).

Sequencing was performed by the Sequencing Team of Queensland Centre for Medical Genomics led by David Miller. The members of the Sequencing team are: Craig Nourse, Ehsan Nourbakhsh, Suzanne Manning, Tim Bruxner, Senel Idrisoglu, Ivon Harliwong, and Angelika Christ.

Pre-processing of sequencing data was performed by the Informatics Team of Queensland Centre for Medical Genomics led by John Pearson. The members of the Informatics team are: Lynn Fink, Felicity Newell, Darrin Taylor, Conrad Leonard, Oliver Holmes, Christina Xu, Matthew Anderson, Scott Wood, Sarah Song and David Wood.

Guidance and advice were provided by the Data Analysis Team of Queensland Centre for Medical Genomics led by Nic Waddell. The members of Data Analysis team are Katia Nones, Ann-Marie Patch, Peter Bailey, Michael Quinn, Karin Kassahn and Nicole Cloonan.

Statistical analysis was designed by Kim-Anh Lê Cao.

I, Kelly Quek, performed the wet lab verification of structural rearrangements, implemented and reviewed the performance of in-house tools, data interpretation, statistical analysis, optimized the detection and quantification of circulating tumour DNA and reviewed patients' disease progression of the analysed serums.

Statement of parts of the thesis submitted to qualify for the award of another degree

None

Acknowledgements

The study of genomic alterations in cancers has accelerated with the advanced of next generation sequencing. In this thesis, I systematically catalogued and characterised somatic variation in a large cohort of 120 pancreatic primary tumours. The work would not have been possible without the help of many people. And so, I would like to take this opportunity to acknowledge and thank everyone for the role they played during my years in QCMG.

My boss, Sean Grimmond for the opportunity to work as a member of QCMG. Nic Waddell, Katia Nones for mentoring and providing me with lots of support.

Deborah Gwynne, for central co-ordination at the Queensland Centre for Medical Genomics.

Ann-Marie Patch, Nicole Cloonan, Kim-Anh Lê Cao, Peter Bailey, Peter Simpson, Mark Ragan, Dave Tang for scientific input and guidance in my projects.

Members of the Sequencing team: David Miller, Craig Nourse, Ehsan Nourbakhsh, Suzanne Manning, Tim Bruxner, Senel Idrisoglu, Ivon Harliwong, and Angelika Christ, for their help in setting up the wet laboratory experiments.

Members of the Informatics team: John Pearson, Lynn Fink, Felicity Newell, Darrin Taylor, Stephen Kazakoff, Conrad Leonard, Oliver Holmes, Nick Waddell, Christina Xu, Matthew Anderson, Scott Wood, Sarah Song and David Wood for their help in designing, maintaining bioinformatics tools and making the sequencing data accessible.

My parents, my family and my friends, for their love, encouragement, and support always.

I was supported by the University of Queensland International Research Tuition Award, University of Queensland Research Scholarship, EMBL PhD travel grant and IMB/UQ travel award throughout the course of my PhD research (2011-2015).

Keywords

Structural variation; Cancer; Next generation sequencing; Genomics; Chromosomal rearrangements breakpoints; Biomarkers; Circulating tumour DNA.

Australian and New Zealand Standard Research Classifications (ANZSRC)

ANZSRC code: 060408, Genomics 60%
ANZSRC code: 111299, Oncology and Carcinogenesis not elsewhere classified 20%
ANZSRC code: 060102, Bioinformatics 20%

Fields of Research (FoR) Classification

FoR code: 0604, Genetics, 50%
FoR code: 1112, Oncology and Carcinogenesis, 50%

Table of Contents

List of Figures and Tables

Main Figures

Supplementary Figures

Main Tables

Supplementary Tables

List of Abbreviations

aCGH	Array Comparative Genomic Hybridization
ALK	Anaplastic Lymphoma Receptor Tyrosine Kinase
AOCS	Australian Ovarian Cancer Study
BeaMing	Beads, Emulsification, Amplification, and Magnetics
BFB	Break-Fusion-Bridge
bp	Base pairs
CA	Carbohydrate Antigen
cfDNA	Cell-free DNA
CLL	Chronic Lymphocytic Leukaemia
CML	Chronic Myeloid Leukaemia
COSMIC	Catalogue of Somatic Mutations In Cancer
CTCs	Circulating Tumour Cells
ctDNA	Circulating Tumour DNA
DDR	DNA Damage Response
DNA	Deoxyribonucleic Acid
dNTP	Deoxynucleotide Triphosphate
dPCR	Digital PCR
DSB	Double Strand Break
GEM	Gemcitabine
GTP	Guanosine Triphosphate
HGMD	Human Gene Mutation Database
HR	Homologous Recombination
ICGC	International Cancer Genome Consortium
kb	Kilo base
LINEs	Long Interspersed Elements
LOH	Loss of Heterozygosity
LTR	Long Terminal Repeat
Mb	Mega base
MMEJ	Microhomology-Mediated End Joining
NHEJ	Non-Homologous End Joining
OESO	Oesophageal adenocarcinoma
PARE	Personalized Analysis of Rearrangement Ends
PCR	Polymerase Chain Reaction
PDAC	Pancreatic Ductal Adenocarcinoma
QCMG	Queensland Centre for Medical Genomics
qPCR	Quantitative Polymerase Chain Reaction
RNA	Ribonucleic Acid
SINEs	Short Interspersed Elements
SV	Structural Variation
TAm-Seq	Tagged-Amplicon Deep Sequencing
TCGA	The Cancer Genome Atlas
TE	Transposon Element
TSG	Tumour Suppressor Gene

1 Chapter 1

Introduction

1.1 Genomic landscape of the cancer genome

Over the last two decades it has become increasingly clear that cancer is initiated by the accumulation of genetic damage in key cellular pathways. This damage ranges from small point mutations affecting a single DNA base to large genomic rearrangements. Collectively, these mutations confer a set of key attributes common to cancer cells (Negrini et al., 2010) (Figure 1-1), also known as 'the hallmarks of cancer'. The disruption of these biological processes by genomic alterations favours cancer progression leading to uncontrolled growth of mutated cells (Negrini et al., 2010; Stratton et al., 2009).

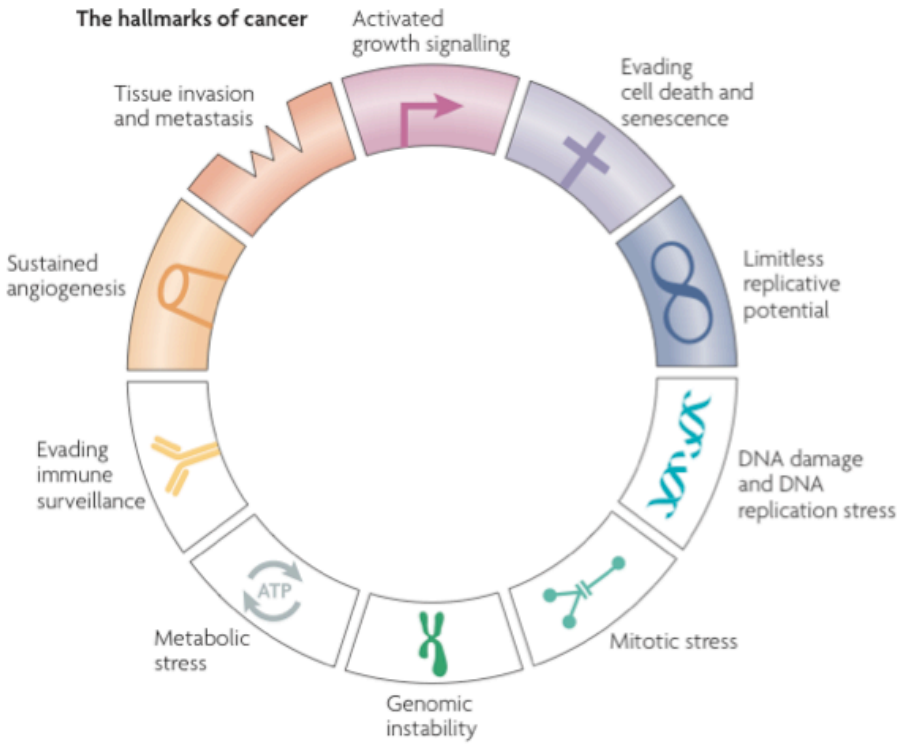


Figure 1-1 The hallmarks of cancer. The hallmarks of cancer describe biological processes related either to functional capabilities or to the presence of stress in cancer cells. Figure adopted from Negrini et al. (Negrini et al., 2010). Image used with permission from Nature Reviews Molecular Cell Biology (License Number: 3506161060186).

Alterations present in the cancer genome typically arise sporadically in response to environmental and lifestyle exposure to mutagens or due to intrinsic replication errors

(somatic mutations). When these changes or mutations occur in key genes or pathways, the cell with these changes may acquire a survival advantage. This cell may begin to divide abnormally. Subsequent cells may acquire new mutations which may result in cancer formation. On the contrary, germline mutations arise less frequently, and individuals are more likely to have genetic predisposition than intrinsic mutations or may have inefficient repair machinery to repair changes caused by exposure to mutagens. Therefore, this increases their risk of developing cancer. These alterations come in many forms including base substitution, small indels (small insertions and deletions), large structural variation (SV; changes to the structure of DNA), copy number variation (gain or loss of DNA sequence), loss of heterozygosity (LOH, loss of one allele) and epigenetic lesion (changes in DNA methylation or histone modification). If these mutations or alterations occur in genic regions, they may have an effect on the function of the gene.

Mutated genes that confer growth advantages and are positively selected in the tissue to promote tumorigenesis are referred as ***driver*** genes. Mutations that do not contribute to tumorigenesis but at some point are carried along in the clonal expansion of cancer cells are referred as ***passenger*** mutations (Stratton et al., 2009). Examples of driver mutations are single base substitution in the *KRAS* oncogene in pancreatic and colon cancers (Edkins et al., 2006) which results in an activated *KRAS* protein, missense mutations in *ARID1A* gene in pancreatic cancer (Biankin et al., 2012), structural rearrangement forming *BCR-ABL* fusion gene in leukaemia (Nowell and Hungerford, 1960), amplification of *EGFR* gene in lung cancers (Cappuzzo et al., 2005) and epigenetic inactivation of p16 in breast cancers (Witcher and Emerson, 2009). In some cases, especially for tumour suppressor genes, both copies of a gene have to be altered (bi-allelic damage) in order to promote a cancer phenotype. This observation was initially described by Knudson as the “two hit” hypothesis. Knudson et al. (Knudson, 1971) conducted a statistical study in retinoblastoma and showed that in the non-hereditary form of retinoblastoma (somatic mutation) both alleles were required to be altered to develop tumour. By contrast, in the inherited form (germline mutation) the first alteration was inherited in the DNA and the second alteration led to tumour development.

In this chapter, I focus on somatic structural rearrangements (also known as structural variation) in human cancers and how such genomic alteration could lead to the development of cancer. I describe the role of chromosomal structural rearrangements in

cancer genomes (a form of genomic instability), different mutational mechanisms and catastrophe events that might be associated with these rearrangements.

1.2 Structural rearrangements

By definition, structural rearrangements are alterations within the genome, which are typically larger than 1 kb (Feuk et al., 2006). Such alterations can result in different types of rearrangements such as deletion, novel sequence insertion, mobile-element insertion, tandem duplication, interspersed duplication, inversion, and translocation (Alkan et al., 2011) (Figure 1-2). **Deletion** refers to the absence of a segment of DNA in a chromosome (Feuk et al., 2006). **Novel sequence insertion and mobile-element insertion** refer to DNA sequence inserted into a given location (Feuk et al., 2006). **Tandem duplication** and **interspersed duplication** involve duplication of a segment of DNA in a chromosome (Feuk et al., 2006). **Inversion** involves breakage inversion and re-insertion of a DNA segment into a chromosome (Feuk et al., 2006). **Translocation** involves exchange or attachment of different DNA segments with no extra or missing genetic information (balanced) or exchange in chromosome material resulting in extra or missing genes (unbalanced) (Feuk et al., 2006).

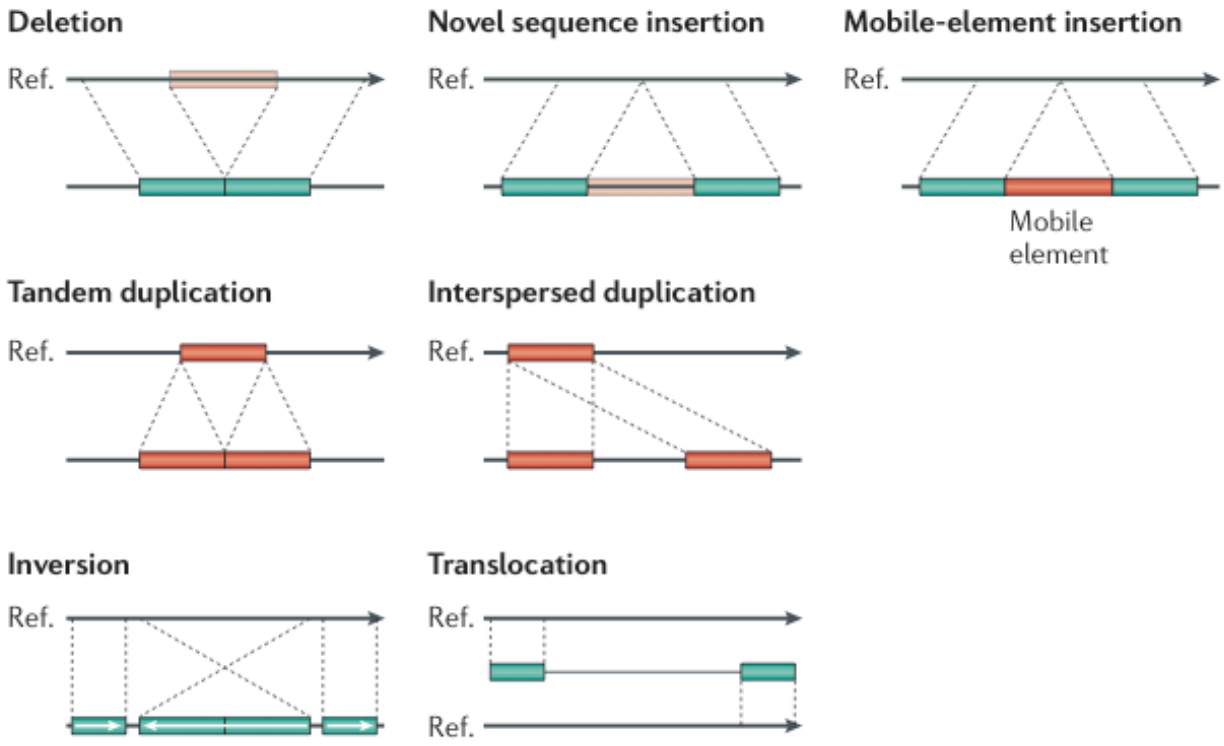


Figure 1-2 Types of structural rearrangements in human cancer. The schematic depicts the different types of structural rearrangements in human cancer. The

advancement in sequencing technologies has led to discovery of small-scale insertions and deletions (<1 kb) referred to as indels as well as large structural rearrangements that can range from 1 kb – 5 Mb (Alkan et al., 2011). Coloured boxes represent read pairs from the tumour and dotted lines represent how the reads map back to the reference genome. Figure adopted from Alkan et al. (Alkan et al., 2011). Image used with permission from Nature Reviews Genetics (License Number: 3506170505158).

1.3 Genomic features that may influence the formation of structural rearrangements in human cancers

A number of genomic features have been shown to influence the occurrence of chromosomal rearrangements in cancers. Cancer-specific rearrangements can be triggered by:

Chromosomal fragile sites. Many studies have demonstrated that fragile sites are hotspots for DNA breakage and there is a strong association between chromosomal breakpoints and the location of fragile sites, suggesting that fragile sites may play a role in the formation of cancer-specific rearrangements (Dillon et al., 2010; Gandhi et al., 2010). For example, genes that span the *FRA3B* fragile sites show increased frequencies of deletions and translocations in cancers (Corbin et al., 2002; Inoue et al., 1997). Furthermore, the DNA sequence at fragile sites is found to be composed of the expansion of CGG repeats or AT-rich sequence (Fungtammasan et al., 2012).

Repetitive regions. Repetitive DNA sequence can be a source of genomic instability in human cancers (Ribeiro et al., 2004). The two main categories of repeated DNA sequences are tandem repeats (e.g. satellite DNA, microsatellite) and interspersed repeats (e.g. DNA transposons, retrotranspsons). Molecular analysis has shown that repetitive DNA sequence is associated with DNA breakage (Argueso et al., 2008). Extending this, repetitive DNA sequences have been reported to be involved in the formation of chromosomal alterations such as amplifications and DNA rearrangements in human cancers (Makela et al., 1992; Ribeiro et al., 2004).

Chromatin modification. Chromatin structures, including epigenetic alterations, have been associated with genomic rearrangements. For example, whole-genome sequencing analysis of human prostate cancer has suggested that DNA rearrangements

are likely to occur in genes that are clustered together within regions of open chromatin (Baca et al., 2013; Berger et al., 2011).

1.4 Consequences of structural rearrangements in human cancers

Most structural variants have been characterised at the molecular level. Functional studies have shown that these large variants can lead to serious consequences and play important roles in initiating cancer. For example:

Disruption of a gene. Large numbers of structural variants directly delete or rearrange sequences of key genes leading to a loss of function. For example, the tumour suppressor genes *BRCA1* and *BRCA2* are commonly disrupted by insertions or deletions resulting in protein truncations (Sadikovic et al., 2008).

Gene copy number alteration. The overall copy number of genes can lead to either a gain of function through increased copy number (amplification) or partial to complete loss of function of a gene. For example, amplifications of oncogenes (such as *KRAS, MYC*) and deletions of tumour suppressor genes (such as *TP53, PTEN*). These events can have functional effects in key pathways involved in tumorigenesis (Leary et al., 2008). Copy number alterations can result in LOH, in which one allele is lost. Gain of the remaining allele can result in copy neutral LOH. Moreover, copy number changes may result in an altered level of expression of cancer-related genes in the tumour cells (Xing et al., 1999).

Formation of fusion genes. This occurs when two formerly separate genes are joined together through genomic rearrangements forming a fusion gene that confers a growth advantages on cancer cells (Stratton et al., 2009). An example is the Philadelphia chromosome in chronic myeloid leukaemia (CML) in which *ABL* gene on chromosome 9 and *BCR* gene on chromosome 22 are joined juxtaposed to create mutant *BCR-ABL1* fusion gene. The chimeric protein encoded by this gene fusion alters tyrosine kinase activity and speeds up cell division (Nowell and Hungerford, 1960). Another example is *EML4-ALK* fusion gene in lung cancer in which chromosome 2 is broken in two locations and the resulting piece of DNA is inverted and re-inserted into the chromosome leading to the fusion of *ALK* and *EML4* genes. This produces a fusion protein that is highly active in ALK kinase activity generating oncogenic effects (Soda et al., 2007).

1.5 Catastrophe events, DNA repair and mutational mechanisms associated with structural rearrangements

Perturbed DNA damage repair enables tumour growth. In a normal cell, the cellular DNA damage response (DDR) machinery is capable of screening for DNA damage, removing altered DNA and restoring correct nucleotide sequence. However, cancer genomes frequently harbour a number of defects in the DDR machinery. The disruption of the DDR machinery allows damaged DNA to survive during cell cycles, thereby increasing the chance of tumorigenesis (Bartek and Lukas).

With the advancement of genomic technologies, catastrophe events and mechanisms such as chromothripsis, kataegis, chromoplexy, and break-fusion-bridge (BFB) were identified to be associated with the formation of somatic rearrangements in cancer genomes. Here, I summarize what is known about these mechanisms and what features of the structural rearrangements may give us clues about the operative mechanisms or events in cancer genomes.

Chromothripsis: This is a mechanism that generates a high number of rearrangements in a complex localized fashion and these complex rearrangements are formed by shattering one or few chromosomal regions in a single event that may lead to rapid cancer development. This phenomenon was first reported in the genome of chronic lymphocytic leukaemia (CLL) sample with a total of 42 rearrangements localized on the long arm of chromosome 4 (Stephens et al., 2011). Chromothripsis is identified by the following key features (Korbel and Campbell, 2013): (1) a high number of rearrangements (clusters) are localized at a particular region within genome; (2) rearrangements show marked oscillation between two or three copy number states; (3) affected regions display segments with retained heterozygosity interspersed with LOH (e.g. regions of copy number 1 show LOH; regions with copy number 2 retain heterozygosity); (4) DNA shattering is typically originated from a specific haplotype; (5) the order and type of rearrangements are random.

Kataegis: This is a localized hypermutation event characterised by clusters of C>T and C>G mutations at TpCpX trinucleotides on the same strand at particular regions along the genome (mutational thunderstorm) (Nik-Zainal et al., 2012). This phenomenon was first reported in the study carried out by Nik-Zainal et al. whereby the authors sequenced 21 breast cancer genomes and observed clusters of mutations at specific regions along

the genome (‘kataegis’) visualised in ‘rainfall plot’. Of the 21 breast cancer, 13 (62%) revealed some extent of kataegis. Extending this study, Alexandrov et al. applied the same approach and analysed 30 different cancer types from whole genome and exome sequencing. Furthermore, it was noted that the regions of kataegis were co-localized with genomic rearrangements (Alexandrov et al., 2013; Nik-Zainal et al., 2012).

Chromoplexy: This is a phenomenon first reported by Berger et al. (Berger et al., 2011) - the authors observed frequent occurrence of large complex chains of rearrangements involving multiple genes in seven prostate tumours. These chain events were generated by the shuffling of broken DNA ends and subsequently re-joined randomly in a novel manner. In a recent study, the analysis of “chromoplexy” was expanded to 57 prostate tumours and a computational method was developed to systematically detect chromoplexy events (Baca et al., 2013). The analysis revealed that many of these chains contained variable number of rearrangements (> 40 rearrangements), which lead to DNA deletions, and fusions genes located near rearrangements breakpoints.

Break-Fusion-Bridge: Overexpression of oncogenes through gene amplification is frequently observed in cancer cells (Hillmer et al., 2011; Lo et al., 2002). This is known that BFB, a DNA replication based mechanism, accounts for such amplification. By its name, the cycles involve the breakage of chromosome followed by fusion and then “bridge” formation. Briefly, BFB begins with the telomere loss, followed by fusions of unprotected chromosomal sister chromatids ends. The fused chromosomes are separated during anaphase and forming a “bridge” and ultimately breaking as the centromeres are pulled in opposite direction. This process repeats for several cell division cycles resulting complex genetic combination with dramatic copy number increases. Furthermore, when the amplified regions harbour oncogenes, this can provide a growth advantage to cancer cells (Bunting and Nussenzweig, 2013).

1.6 Detection of structural variation

The combination of experimental studies and computational strategies has revealed extensive presence of large structural variants in cancer genomes (Feuk et al., 2006; Iafrate et al., 2004; Levy et al., 2007; Medvedev et al., 2009; Tuzun et al., 2005). Here, I describe a brief history of molecular biology approaches and genomic technologies to detect structural rearrangements.

Structural rearrangements were first investigated by karyotyping. This technique examines chromosomes under a light microscope and is limited to the physical characteristic of chromosomes (such as length, size difference, and position of centromeres) (Heim and Mitelman, 1992; Warburton, 1991). Subsequently, various cytogenetic banding techniques such as G-banding, spectral karyotyping, and fluorescence in-situ hybridization were used to detect structural rearrangements including translocations, deletions, duplications, insertions and inversions (McNeil and Ried, 2000). The resolution of these later techniques is typically 1-2 Mb, and hence, they are not able to fully describe highly complex rearrangements. Simultaneously, the development of genome-wide array-based (e.g. array comparative genomic hybridization - aCGH (Pinkel et al., 1998) and targeted PCR based approaches (e.g. real-time quantitative PCR - qPCR) have allowed systemically screening of sub-microscopic unbalanced structural rearrangements with varying degrees of resolution (Dhami et al., 2005; Neill et al., 2010; Tagawa et al., 2004). For example, aCGH compares two labelled samples (test and reference) to a set of hybridization targets and qPCR screens for the targeted region of the genome. However, these techniques are limited to detect gain or loss of DNA material (such as deletions and amplifications) and do not detect event types which do not result in copy number changes (such as inversions).

More recent, the introduction of next generation sequencing technologies has expedited the interrogation process of structural variation and overcome some of the limitations encountered by the former techniques and also enabled us to identify the breakpoints of different event types at single nucleotide resolution (Kidd et al., 2008; McKernan et al., 2009; Shendure and Ji, 2008).

1.7 Next generation sequencing

Sequencing platforms

Automated DNA sequencing has stemmed from the human genome project more than a decade ago. The growing demand of high throughput sequencing led to the development of next generation sequencing platforms in recent years. The advanced technologies have impacted the field of cancer genomics by dramatically increasing the pace of discovery of alterations in cancers. Here, I have grouped the next generation sequencing platforms by their sequencing scale throughput and summarized them in Table 1-1.

Table 1-1 Comparison of next generation sequencing platforms

	Small-scale sequencing	
	Ion Torrent PGM (Life Technologies)*	*MiSeq (Illumina)*
Application	Targeted DNA/RNA sequencing, copy number analysis, small RNA sequencing	Small genome, amplicon, and targeted gene sequencing.
Output per run	60-100 Mb	0.3-15 Gb
Read length	400 bp	2 × 300 bp
Run time	3.7 hr	5-65 hr

*Based on Ion 314™ Chip v2

	Medium-scale sequencing		
	Ion Proton™ System (Life Technologies)*	*NextSeq 500 (Illumina)*	*PACBIO RS II - SMRT sequencing (Pacific Bioscience)*
Application	Whole genome, exome, transcriptome, targeted gene sequencing and more	Whole genome, exome, transcriptome sequencing and more.	De novo assembly, targeted sequencing, base modification, metagenomics
Output per run	~10 Gb	20-120 Gb	100-150 Mb/SMRT cell
Read length	200 bp	2 x 150 bp	> 8,000 bp (C3 chemistry)
Run time	2-4 hr	15-30 hr	90 min

*Based on Proton I

	Large-scale sequencing	**X Large-scale sequencing**
	HiSeq 2500 (Illumina)	*HiSeq X Ten (Illumina)*
Application	Whole genome, exome, transcriptome sequencing and more	Whole genome sequencing
Output per run	10-1000 Gb	1.6-1.8 Tb
Read length	2 × 150 bp, 2 × 125 bp	2 × 150 bp
Run time	7hr – 6 days	< 3 days

Gb – Giga base, Mb – Mega base, hr – hour

To attain a comprehensive view of disease specific mutations in cancer genomes in particular structural rearrangements, whole genome sequencing is required to compare tumour and matched normal DNA of a patient to identify novel somatic structural rearrangements. Today, the most popular platform used is the Illumina HiSeq 2500.

Briefly, Illumina platforms adopt sequencing by synthesis chemistry. Fragments of DNA are immobilized onto the flow cell surface. The fragments are exposed to DNA polymerase to synthesize the complementary strand. The double-stranded DNA is denatured and extended, priming occurs when the free end of a ligated fragment “bridges” over to a complementary oligo in the flow cell surface. The amplification process repeats

and occurs simultaneously forming millions of clusters across the flow cell surface. Sequencing begins with the incorporation of fluorescently-labelled deoxynucleotide triphosphate (dNTP) to the sequences of the nucleic acid chain during each cycle. For each dNTP incorporation, the fluorescent dye emits a signal to identify the base.

As for the small-scale benchtop sequencers (such as Ion Torrent PGM and Illumina MiSeq), they are normally used in targeted sequencing applications such validation of somatic mutations or a predefined region of the genome. MiSeq uses similar chemistry to Illumina HiSeq 2500. Ion Torrent technology uses different sequencing chemistry. It uses semiconductor chip technology to detect hydrogen ions produced during DNA replications in real time. The semiconductor chip contains millions of wells capturing single stranded DNA molecule for sequencing. The sequencing process begins with DNA fragments attached to the surface of bead particles and then clonally amplified. The templated bead deposited into the wells on the semiconductor chip and followed by each of the four nucleotides is orderly introduced. When the nucleotide is incorporated into a single strand of DNA, a hydrogen ion is released. The ion senor records the change in pH of the solution indicating that the nucleotide has incorporated.

Overall, next generation sequencing technologies enable sequencing in a massive parallel manner allowing large number of samples to be sequence at a time. The use of such technologies facilitates efficient and economic genome wide readout on molecular level. The data obtained from next generation sequencing has much higher resolution, which can identify somatic structural rearrangements at nucleotide level.

1.8 Large integrative cancer genome programs and databases

In order to catalogue mutations in a broad range of cancer types, two large international consortiums have been established to coordinate the catalogues of mutations in a large number of cancer genomes around the world. They are the Cancer Genome Atlas (2011) (TCGA) and the International Cancer Genome Consortium (2010) (ICGC). The goals of these programs are to obtain comprehensive catalogue of somatic mutations in human tumours and cancer cell lines, identify repertoire of cancer genes as well as integrating somatic mutations with epigenomic and transcriptomic studies, functional and pathways analysis. Additionally, the collection of these somatic mutations is stored in public databases such as the ICGC data portal, COSMIC (**C**atalogue **O**f **S**omatic **M**utations In **C**ancer) or HGMD (**H**uman **G**ene **M**utation **D**atabase).

The Queensland Centre for Medical Genomics (QCMG) led by Prof. Sean Grimmond is part of ICGC. Australia's contribution to ICGC is to sequence and analyse pancreatic and ovarian cancers enabling a deeper understanding of the mechanisms that lead to genomic instability and ultimately cancer. The Centre is committed to deliver and enable personalized medicine for the state of Queensland and Australia (www.qcmg.org). Collection of all tumour samples was coordinated via our collaborators at the Garvan Institute (Prof. A. Biankin – Pancreatic project) and Peter MacCallum Cancer centre (Prof. David Bowtell – Ovarian project). A total of 392 primary pancreatic ductal adenocarcinomas, 100 pancreatic neuroendocrine tumours and 93 ovarian tumours underwent genome sequencing analysis at QCMG and were submitted to the ICGC data portal (www.icgc.org | Numbers correct on January 2015).

1.9 Pancreatic cancer

Epidemiology of pancreatic cancer

Pancreatic cancer is an aggressive disease with a poor prognosis and is the 4th largest cause of cancer death in Western countries. Each year over 200,000 people are diagnosed with the disease worldwide. Australia has one of the highest incidence rates of pancreatic cancer in the world (Figure 1-3). The mortality rate of pancreatic cancer is also one of the highest among the major cancers and the survival rate has not improved over the last 40 years with a median survival of 6 months. Only 5% of patients diagnosed with pancreatic cancer survive more than 5 years (Jemal et al., 2010).

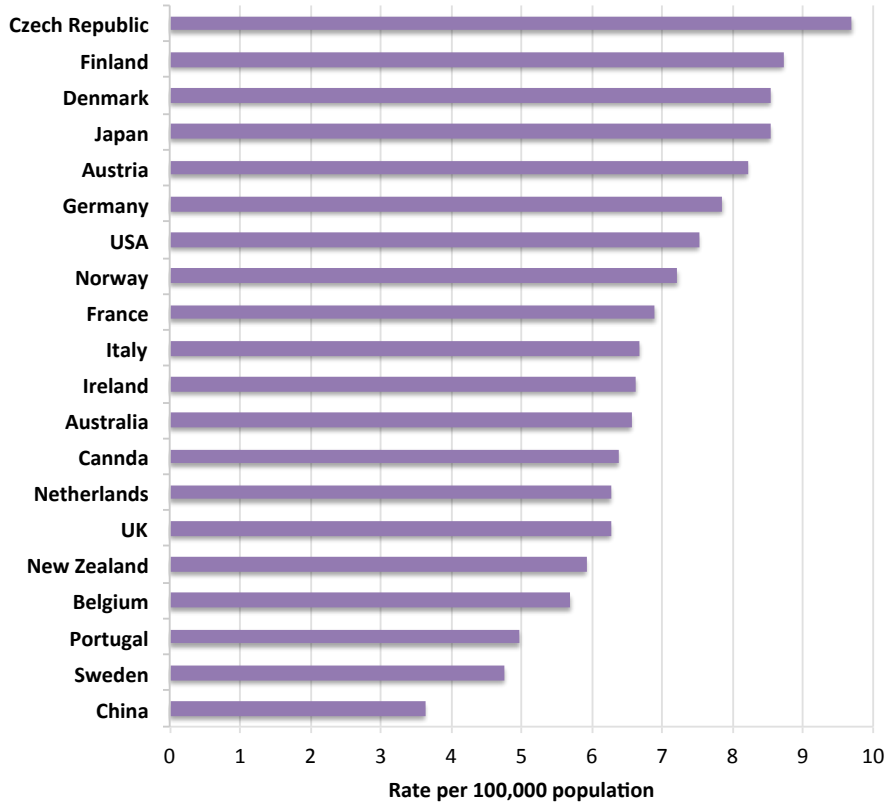


Figure 1-3 World incidence rates of pancreatic cancer. The data was collected from Cancer Research UK in year 2012. This section was last reviewed and updated on 11 June 2014.

There are two main types of pancreatic tumours, exocrine (derived from enzyme producing cells) and endocrine (derived from hormone producing cells). The most common form of pancreatic tumour is exocrine type, which account for more than 95% of all pancreatic tumours. Pancreatic ductal adenocarcinoma (PDAC) is the most common exocrine tumour, making up 90% of all exocrine tumours. The remaining is made up by acinar cell carcinoma, intraductal papillary mucinous neoplasm, mucinous cystic neoplasm, pancreatoblasma, serous cystadenocarcinoma and solid psuedopaillary neoplasm.

On the other hand, pancreatic endocrine tumours make up approximately 5% of all pancreatic cancers and consist of two subtypes: neuroendocrine tumours and islet cell tumours. In this project, the research focuses on exocrine PDAC.

1.9.1 Treatment options

At present, there is no clear screening method to detect pancreatic cancer at an early stage. Due to the lack of early warning symptoms, patients are normally diagnosed at late

stages and the disease has metastasized upon diagnosis. Large tumours can be detected by computed tomography scan, magnetic resonance imaging and endoscopic ultrasound (Klapman and Malafa, 2008).

The only curative treatment is surgery to resect the tumour, followed by chemotherapy. However, only 7-20% of patients present with operable tumour. The remaining non-resectable patients present late stages of the disease are offered palliative chemotherapy that usually involves the chemotherapeutic agent Gemcitabine (GEM) as first line therapy (Saif, 2006). GEM response is only seen in 20-30% of patients and chemoresistance is rapidly acquired. The failure of current treatment regime indicates that there is a need to increase our understanding of this disease and develop alternative treatments for the pancreatic cancer patients.

1.9.2 Molecular characteristic of pancreatic cancer

It is estimated that 90% of pancreatic tumours arise from precursor lesions of DNA during life and these damages often arise sporadically. The exact mechanism of pancreatic cancer development is not yet fully understood but studies have identified key genetic mutations and signalling pathways associate with its tumorigenesis (Biankin et al., 2012; Campbell et al., 2010; Jones et al., 2008; Wang et al., 2012).

The most common genetic lesions that are found in most PDAC are mutations of $p16^{INK4A}$*(CDKN2A)*, *TP53* and *SMAD4(DPC4)* tumour suppressor genes (TSG) and the *KRAS* oncogene. Inactivation of TSG or activating mutations of oncogenes could create a "domino effect", where loss of tumour suppressor genes or activating mutations of oncogenes can affect several cellular pathways that regulate cell-cycle, cell survival, invasion and metastases (Raimondi et al., 2009).

1.9.3 TSG in PDAC

$p16^{INK4A}$ tumour suppressor protein is inactivated in more than 90% of sporadic pancreatic cancer (Caldas et al., 1994). It is located on chromosome 9p21, encoded by the *CDKN2A* gene. *CDKN2A* is an inhibitor of cyclin-dependent kinase 4, when active, it triggers retinoblastoma phosphorylation to induce cell cycle arrest in G1 and G2 phases (Schutte et al., 1997; Sherr, 1996). The mechanisms of *CDKN2A* inactivation include intragenic mutation, deletion and hypermethylation of p16 (Schutte et al., 1997). Inactivating

mutations of *TP53* genes are also found in 50-75% of pancreatic cancer. p53 has multiple tumour suppressing functions and play a crucial role in cell cycle progression by inducing growth arrest or apoptosis in normal cell when the DNA is damaged (Kalthoff et al., 1993; Levine, 1997).

Another TSG frequently mutated in pancreatic ductal adenocarcinoma is *SMAD4*(*DPC4*), which is a member of *SMAD* family of signal transduction proteins which is lost in 50% of pancreatic cancer (Hahn et al., 1996). It acts as a co-activator and mediator in transforming growth factor (TGF-β) signalling pathway by forming complexes with other *SMAD* proteins, then translocated into the nucleus and regulates the expression of target genes (Heldin et al., 1997). In the context of cancer, the inactivating mutations of *SMAD4*(*DPC4*) may disrupt TGF-β signalling pathway and then up-regulate the expression of cancer-associated genes to facilitate cancer tumorigenesis.

1.9.4 Oncogenes in PDAC

Mutations of *KRAS* oncogene are present in more than 90% of pancreatic cancers (Pellegata et al., 1994). This gene is mutated in different positions in its sequence codon, for example, codon 12, 13 or 61. The *KRAS* gene encodes a protein with GTPase activity that mediates signal transduction pathway regulating cell cycle progression and cell proliferation. The single amino acid substitution at codon 12, 13 or 61 leads to perpetual activation of signalling with the loss of GTPase activity locking KRAS into an active state.

1.9.5 Mutational landscapes and pathways in pancreatic cancer

Sequencing efforts have revealed mutational landscape of pancreatic cancer (Biankin et al., 2012; Jones et al., 2008; Wang et al., 2012). The complexity of the genetic alterations in PDAC was first described in the study conducted by Jones et al. (Jones et al., 2008). Genomic analysis of more than 900 genes across 24 PDAC xenograft and cell lines have shown that these mutations lead to recurrent perturbation of 12 core signalling pathways which frequently altered (67-100%) among the cohort of 24 patients (Figure 1-4).

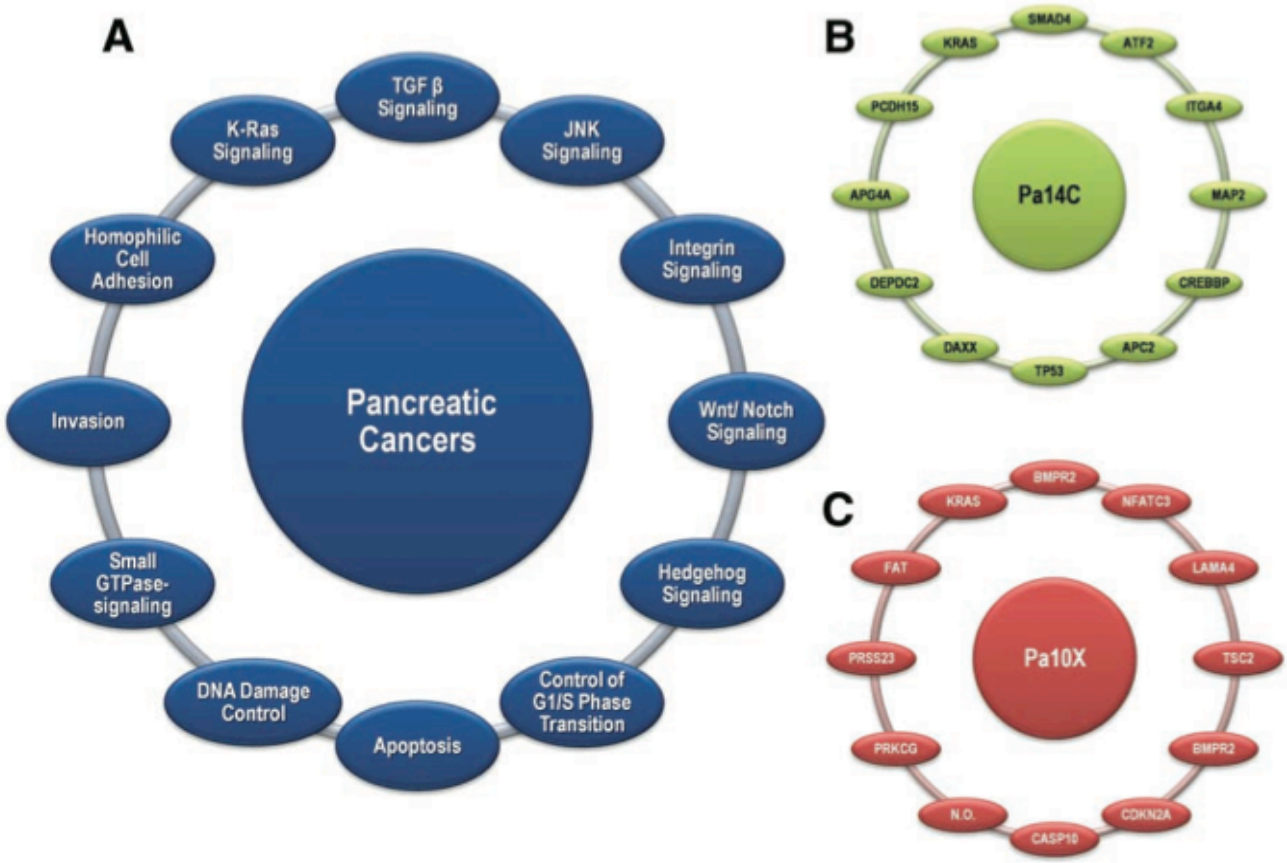


Figure 1-4 Signalling pathways and processes. (a) The 12 core signalling pathways and processes genetically altered in most pancreatic cancers. (b) and (c) are the two pancreatic cancers studied in Jones et al (Jones et al., 2008). The positions around the circles in (b) and (c) correspond to the pathways and processes in (a). This figure has illustrated that the common signalling pathways share a number of genes, for instance, the mutation in BMPR2 have disrupted both SMAD4 and Hedgehog signalling pathways in Pa10X (Jones et al., 2008). Image used with permission from Science (License Number: 3527851309934).

To date, the largest study of mutations in PDAC involved exome sequencing and CNV analysis of 142 PDAC primary tumours (Biankin et al., 2012). The vast majority of mutations identified by Biankin et al. are somatic in nature (such as missense, nonsense, splice site, insertion/deletion, non-silent and silent). The study has confirmed the importance of known mutations in PDAC such as *KRAS, TP53, CDKN2A, SMAD4, MLL3, TGFBR2, ARID1A and SF3B1* and uncovered novel significant mutated genes such as *EPC1* and *ARID2* (involved in chromatin modification), *ATM* (implication in BRCA-mediated DNA damage repair mechanisms) and many more which occur at low frequency. Pathway analysis also revealed that these novel genes were strongly associated with axon

guidance pathways. Additionally, recurrent mutations (*SLIT2 and ROBO2*) of axon guidance pathways were identified in 20% of the patients suggesting that this pathway might play a role in the pathophysiology of PDAC disease.

1.10 Pan-cancer analysis

Thousands of tumours of many types have been sequenced by TCGA and ICGC for the discovery of mutations to deepen our understanding of the nature of cancer genomes. Through these studies, we learnt that cancer genomes display unique spectrum of mutations within each cancer type. However, it is believed that a set of mutations (i.e. driver mutations) may involve in certain pathways and/or mechanisms that could be shared by more than one cancer type. As such, TCGA Research Network set up pan-cancer initiative studies to identify possible common patterns shared by different cancer types.

In 2013, the TCGA Research Network released a series of papers to compare the differences and similarities between tumours originated from different tissues (http://www.nature.com/ng/focus/tcga/index.html). One example of these studies was conducted by Kandoth et al. analysing 3,821 primary tumours across 12 cancer types (Kandoth et al., 2013). More than 600,000 somatic mutations were identified including missense, silent, nonsense, splice site, non-coding DNA, 'non-stop', and indels. Detailed analysis of the integrated data revealed 127 significantly mutated genes from known signalling and enzymatic processes across cancers and 66 of them were driver mutations. Furthermore, they learnt that the combination of driver mutations varies within individual patients. Hence, such studies provide new insights of the wide spectrum mutations in cancer genomes and their molecular composition.

Taken together, the efforts of these emerging studies looking across different cancer types and/or individual patient tumour enable us to fully exploit the information of cancer genomics to (1) improve the development of prognostic/diagnostic biomarkers, (2) develop targeted therapeutic intervention, and (3) provide alternative directions for drug development and search more appropriate "druggable" target.

1.11 Summary

In summary, the study of genomic landscape (such as genomic instability and mutational mechanisms) using next generation sequencing holds great promise for restructuring the

way in which we treat cancer. In this thesis, I study somatic structural rearrangements in pancreatic cancer genomes using next generation sequencing. The study of somatic rearrangements in pancreatic cancer genomes is divided into 3 different aspects: genomics technologies and computational methods to detect and verify somatic rearrangements, genomics analysis of somatic rearrangements and DNA patterns in the region of breakpoints, and the utility of somatic rearrangements in personalized genomics medicine. Figure 1-5 detailed what will be investigated in the following chapters of this thesis.

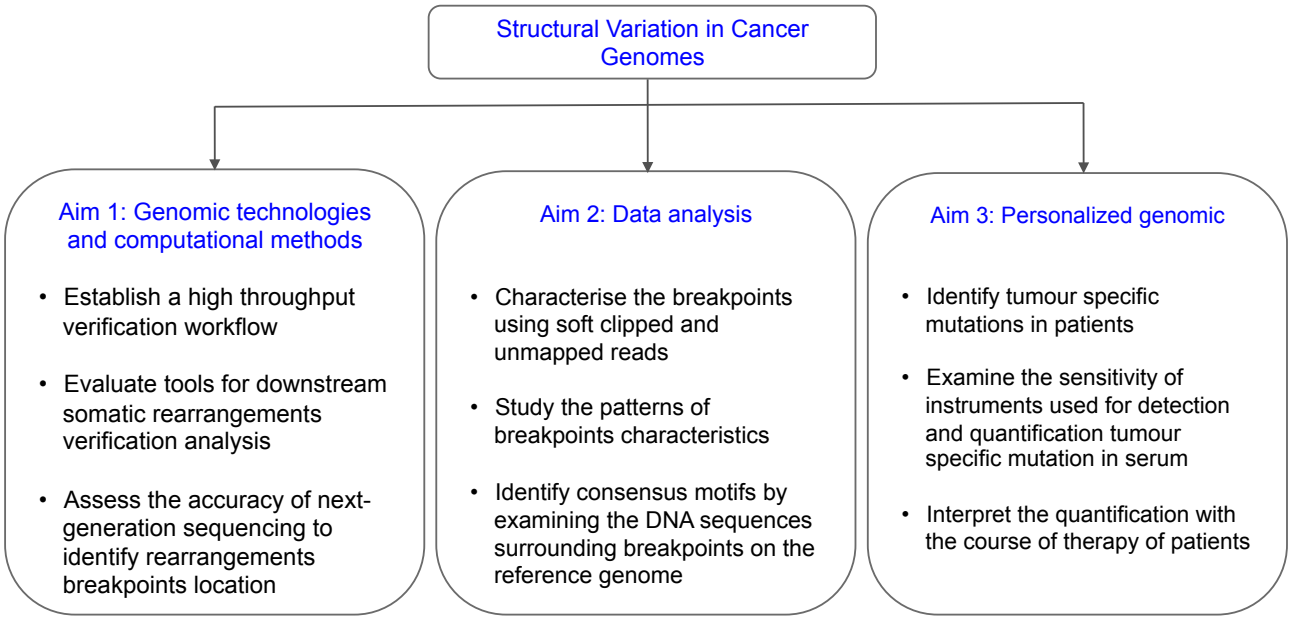


Figure 1-5 An overview framework of the thesis. Aim 1 - Chapter 2: Genomics technologies and computational methods. This project presents a high throughput workflow, which can increase the verification rate of somatic rearrangements in cancer genomes. Aim 2 - Chapter 3: Data analysis. The project focuses on the characterization of somatic rearrangements and interpretation of data at large-scale to elucidate the potential mechanisms that contribute to the formation of somatic rearrangements. Aim 3 - Chapter 4: Personalized genomic. The project aims to demonstrate the utility of tumour specific mutations as biomarker for pancreatic cancer diagnosis and therapeutic strategies.

1.12 Hypotheses

Cancer is a genetic disease, which often associates with genomic instability and DNA damage. It is hypothesized that the patterns of somatic breakpoints and the DNA sequence surrounding the breakpoints junctions can identify the DNA repair mechanisms underlying the formation of structural variation in pancreatic cancer genomes.

We further hypothesize that tumour specific mutations can be used as candidate biomarkers in pancreatic cancer. Tumour specific rearrangements and mutations in pancreatic cancer can be detected in cell-free circulating DNA of patients. These rearrangements and mutations can be used as personalized biomarkers to determine tumour burden and monitor patient response to therapy.

1.13 Aims

To address these hypotheses, the aims are to:

1. Develop an approach utilizing next generation sequencing to rapidly capture and verify somatic chromosomal rearrangements in the genome of primary tumours and establish a high throughput genomic and computational workflow

2. Identify the exact location of DNA breakpoints and characterise the breakpoints region that may indicate potential mechanisms that contribute to the formation of somatic rearrangements.

3. Identify personalized DNA-based biomarkers for pancreatic cancer and assess whether they can be used to monitor tumour burden and response to chemotherapy

1.14 Significance

The work presented here is to explore genomic data of pancreatic cancer. The aim of this study is to catalogue and verify large numbers of somatic chromosomal rearrangements detected by next generation sequencing. The resulting high-resolution catalogue of verified rearrangements and the sequence context of breakpoints together with other somatic mutations would provide clues of how somatic variants implicate in cancer establishment and maintenance. And lastly, we will evaluate the utility of verified somatic rearrangements and mutations as biomarkers in the serum or plasma of pancreatic cancer patients in the hope to develop an alternative approach to trace the course of the disease also during cancer treatment.

More importantly, the work could show the potential of using next generation sequencing technology as a tool to understand different aspects cancer genomes.

2 Chapter 2

A workflow to increase verification rate of somatic chromosomal rearrangements using high throughput next generation sequencing

2.1 Introduction

The ICGC project aims to catalogue the complete repertoire of somatic mutations, identify drivers of mechanism of cancer development and progression and improve therapy options to ultimately benefit patient's outcome. QCMG is part of the ICGC (International Cancer Genome Consortium) and in 2014, we completed the sequencing and analysis of 392 primary pancreatic ductal adenocarcinomas, 100 pancreatic neuroendocrine tumours and 93 ovarian tumours. During this process, the laboratory not only generated and analysed a large volume of sequencing data but also developed several bioinformatics tools to facilitate the genomics analysis of these tumours (Kassahn et al., 2013; Quek et al., 2014; Song et al., 2012). The sequence data was analysed by the QCMG to determine somatic base pair substitutions, indels, copy number changes and chromosomal structural rearrangements. Within this large initiative, my first results chapter specifically aimed to detect and verify somatic rearrangements obtained from whole genome sequencing data and create a comprehensive catalogue of high quality structural mutations.

2.1.1 qSV

There are five main methods to detect structural rearrangements: (1) discordant read pair, (2) soft clipping, (3) spilt read, (4) read depth and (5) sequence assembly (Alkan et al., 2011; Wang et al., 2011). The first 4 methods utilize 'comparison-versus-reference' strategy, whereby next generation sequencing data is aligned to a known reference genome and then the genomic locations of potential structural rearrangement are estimated. The 5th method is based on sequence assembly, which joins DNA sequences with similarity and constructs into a larger census contig. At QCMG, we developed a tool named qSV to identify somatic rearrangements using 3 methods - discordant read pair, soft clipping and split read (Figure 2-1) (Patch*, Newell* and Quek* et al. – manuscript in

preparation). The brief descriptions of the 3 methods incorporated in the detection tool are as follows:

*1) Discordant read pair***.** Paired end sequencing data consists of two sequence reads, which should align to the genome at a known distance apart. The discordant read pair method assesses the distance and orientation of the two sequence reads and identifies read pairs in which distance and orientation of the read pairs are inconsistent with the reference genome. For example, deletion in which read pairs are mapped in the correct orientation but too far apart, or insertion in which read pairs mapped too close and inversion or tandem duplication where read pairs are in incorrect orientation (Alkan et al., 2011). Other computational tools which use discordant read pair method include BreakDancer (Chen et al., 2009), DELLY (Rausch et al., 2012), VariationHunter (Hormozdiari et al., 2009) and PEMer (Korbel et al., 2009).

*2) Soft clipping***.** Soft clipping occurs if part of a read maps well, while the other end of the read does not align. The unaligned sequence is hidden but retained in the BAM file (known as soft clipping). In the case of breakpoints, the unaligned part of the sequence read will align to the other side of the breakpoint. This soft clipping method identifies clusters of reads, which have been soft clipped to detect breakpoints and then directly find the exact breakpoint of the structural rearrangements. Computational tools, which use the soft clipping method, are CREST (Wang et al., 2011) and Socrates (Schroder et al., 2014).

*3) Split read***.** The split read method can directly detect the breakpoints of structural rearrangements by mapping the read sequences to the reference genome with gap alignment. qSV uses split read logic – it performs a round of de novo assembly on all unmapped, abnormally mapped and clipped bases from reads, whose pair is mapped within +/- 1 Kb of putative breakpoint by qAssemble. The longest assembled contig is subjected to local alignment on the reference genome carried out using BLAT (Kent, 2002). This process increases breakpoint resolution as the larger assembled product is able to capture more of the complexity of a locus than soft clipping signature alone. Hence, the assembly of a longer contig enable the discovery of deletions and small insertions. Microhomology and non-template sequenced are calculated and refined for soft clipping breakpoints and generated where possible for events with discordant mapped pairs evidence only.

Computational tools such as DELLY (Rausch et al., 2012), PRISM (Jiang et al., 2012), and Pindel (Grabherr et al., 2011) use split read algorithm to uncover structural rearrangements in human genomes.

Evidence of all the 3 methods increases the confidence that an event is real (Alkan et al., 2011). This strategy clusters different rearrangements with the same characteristic to support a potential structural rearrangement (Figure 2-1).

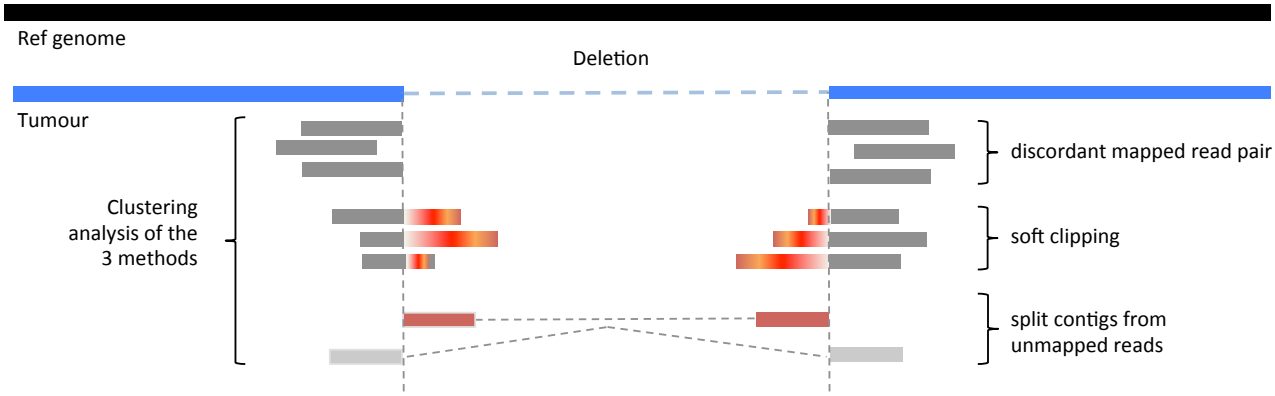


Figure 2-1 An illustration of qSV analysis. The tool uses discordant read pair, soft clipping and split read methods followed by clustering analysis to build evidence that an event is real.

To achieve high quality identification of somatic mutations, it is essential for genomic variants in the tumours to be manually curated and/or verified experimentally. Conventionally, the standard workflow used to verify somatic rearrangements in the literatures employs Sanger sequencing. Despite the strength of Sanger sequencing, it requires extensive labour input and time to verify a large number of events, especially when cleaning up individual PCR products and preparing individual samples for Sanger sequencing. Each amplicon needs to be quantified and the sequencing reaction is prepared individually as the amount of input DNA depends on the size of the amplicon. In this chapter, I present a high throughput workflow, which uses benchtop next generation sequencers to verify many events in one experiment.

2.2 Results and discussion (Presented as form of manuscript)

Quek et al. 2014. BioTechniques *57*:3-38. A workflow to increase verification rate of chromosomal structural rearrangements using high throughput next generation sequencing (Quek et al., 2014).

Essentially, a high throughput workflow utilizing pooling strategy of amplicons combined with benchtop sequencing and standard bioinformatics techniques was established to increase the efficiency of somatic rearrangements verification. The proposed workflow facilitates the verification at base resolution of hundreds of breakpoints in a single experiment. The workflow was essential for tuning/developing the upstream analysis carried out by QCMG sequencing and bioinformatics team. We compared sequences and breakpoints of verified somatic rearrangements between the conventional and high throughput workflow. The results showed that next generation sequencing methods are comparable to conventional Sanger sequencing. The identified breakpoints obtained from next generation sequencing methods were highly accurate and reproducible. Furthermore, the proposed workflow allows hundreds of events to be processed in a shorter time frame compared to the conventional workflow.

All figures and supplementary data presented in this chapter include

Figure 2-2 Conventional and high throughput workflows for the verification of somatic rearrangements and identification of breakpoints.

Figure 2-3 Verification of somatic structural rearrangements from a highly rearranged cancer genome.

Figure 2-4 The ability of different sequencing approaches to resolve the sequence of the breakpoints.

Table 2-1 Summary of primers designed and verification rate for pancreatic cancer genome using qAmplicon and PCR analysis.

Supplementary Figure 2-1 PCR verification of candidate somatic rearrangements using short amplicon primers.

Supplementary Figure 2-2 Classification of PCR verification of candidate somatic rearrangements.

Supplementary Figure 2-3 Breakdown of verification results for 311 candidate somatic events by both short and long amplicons.

Supplementary Figure 2-4 An illustration of the gapped alignment of the sequence reads taken from an intra-chromosomal rearrangements.

Supplementary Table 2-1 List of structural variants called by qSV, validated by PCR and re-sequencing. (Data can be downloaded from http://goo.gl/5Le9KQ)

Supplementary Table 2-2 Summary of events and verification rates for pancreatic cancer genome using HiSeq. (Data can be downloaded from http://goo.gl/5Le9KQ)

A workflow to increase verification rate of chromosomal structural rearrangements using high-throughput next-generation sequencing

Kelly Quek[1], Katia Nones[1], Ann-Marie Patch[1], J. Lynn Fink[1], Felicity Newell[1], Nicole Cloonan[1], David Miller[1], Muhammad Z. H. Fadlullah[1], Karin Kassahn[1], Angelika N. Christ [1], Timothy J. C. Bruxner[1], Suzanne Manning[1], Ivon Harliwong[1], Senel Idrisoglu[1], Craig Nourse[1], Ehsan Nourbakhsh[1], Shivangi Wani[1], Anita Steptoe[1], Matthew Anderson[1], Oliver Holmes[1], Conrad Leonard[1], Darrin Taylor[1], Scott Wood[1], Qinying Xu[1], Australian Pancreatic Cancer Genome Initiative[2], Peter Wilson[1], Andrew V. Biankin[3,4,5,6], John V. Pearson[1], Nic Waddell[1], and Sean M. Grimmond[1,6]

[1]Queensland Centre for Medical Genomics, Institute for Molecular Bioscience, University of Queensland, St Lucia, Brisbane, QLD, Australia, [2]Australian Pancreatic Cancer Genome Initiative (for full list of contributors, see www.pancreaticcancer.net.au/apgi/collaborators), [3]The Kinghorn Cancer Centre, Cancer Research Program, Garvan Institute of Medical Research, Sydney, NSW, Australia, [4]Department of Surgery, Bankstown Hospital, Sydney, NSW, Australia, [5]South Western Sydney Clinical School, Faculty of Medicine, University of NSW, Liverpool, NSW, Australia, and [6]Wolfson Wohl Cancer Research Centre, Institute for Cancer Sciences, University of Glasgow, Glasgow, Scotland, United Kingdom

BioTechniques 57:31-38 (July 2014) doi 10.2144/000114189
Keywords: next-generation sequencing; cancer; chromosome breakpoints; structural variation; verification; high-throughput

Supplementary material for this article is available at www.BioTechniques.com/article/114189.

Somatic rearrangements, which are commonly found in human cancer genomes, contribute to the progression and maintenance of cancers. Conventionally, the verification of somatic rearrangements comprises many manual steps and Sanger sequencing. This is labor intensive when verifying a large number of rearrangements in a large cohort. To increase the verification throughput, we devised a high-throughput workflow that utilizes benchtop next-generation sequencing and in-house bioinformatics tools to link the laboratory processes. In the proposed workflow, primers are automatically designed. PCR and an optional gel electrophoresis step to confirm the somatic nature of the rearrangements are performed. PCR products of somatic events are pooled for Ion Torrent PGM and/or Illumina MiSeq sequencing, the resulting sequence reads are assembled into consensus contigs by a consensus assembler, and an automated BLAT is used to resolve the breakpoints to base level. We compared sequences and breakpoints of verified somatic rearrangements between the conventional and high-throughput workflow. The results showed that next-generation sequencing methods are comparable to conventional Sanger sequencing. The identified breakpoints obtained from next-generation sequencing methods were highly accurate and reproducible. Furthermore, the proposed workflow allows hundreds of events to be processed in a shorter time frame compared with the conventional workflow.

Cancer is the result of the accumulation of genetic damage in key genes and pathways, which ultimately leads to uncontrolled growth of mutated cells (1,2). This damage ranges from small point mutations to large chromosome structural rearrangements. Structural rearrangements include deletions, insertions, tandem duplications, inversions, and translocations. Many cancer genomes carry tens to hundreds of structural rearrangements that may

METHOD SUMMARY

To increase the efficiency of verification of chromosomal structural rearrangements, we present a high-throughput workflow utilizing an amplicon pooling strategy combined with benchtop sequencing and standard bioinformatics techniques that facilitates the examination of more than 300 breakpoints, resulting in the identification of breakpoints in more than 80% of events at single base resolution.

have functional consequences such as disruption of tumor suppressor genes, activation of oncogenes, gene copy number alteration, and/or formation of new fusion genes (3–6).

Traditionally, structural rearrangements have been investigated by cytogenetic methods or array-based approaches. However, these techniques are low resolution, low-throughput, and do not fully capture all types of rearrangements. With the advent of next-generation sequencing, structural rearrangements can be identified to base level resolution using whole genome paired sequencing methods. Sequencing data are mapped to a reference genome, and potential structural rearrangements are identified using approaches such as discordant read pairs (7), split read (8), or soft clipping (9). Two large international consortia, the International Cancer Genomic Consortium (ICGC) and The Cancer Genome Atlas (TCGA), have been established to interrogate different cancer types. Both generate high quality comprehensive catalogs of genomic abnormalities (somatic mutations, abnormal expression of genes, and epigenetic modifications) using next-generation sequencing to better understand the molecular pathophysiology of cancers (10,11). To achieve high quality data, it is essential for genomic abnormalities in the tumors to be curated and verified for the catalog.

With respect to structural rearrangements, the verification workflow in previous studies utilizes manual primer design of each rearrangement, PCR, gel electrophoresis, and Sanger sequencing (12–14). This approach is low-throughput and requires extensive labor input, especially when verifying a large number of rearrangements; consequently it can take weeks to complete the entire verification process.

To increase verification throughput, we have devised a high-throughput workflow that utilizes benchtop next-generation sequencing (e.g., Ion Torrent PGM or Illumina MiSeq) and bioinformatics tools for primer design, de novo assembly of sequencing data, and BLAT to identify the DNA sequence of breakpoints for hundreds of structural rearrangements. As proof of principle, we employed this high-throughput workflow to verify structural variants in a highly complex cancer genome. We assessed the impact of the verification rate when primers were placed at a greater distance from the breakpoints and also examined the performance of the two benchtop next-generation sequencing approaches to identify breakpoints and compare with those obtained by Sanger sequencing. We conclude that the high-throughput verification workflow incorporating next-generation sequencing methods is comparable to the conventional methods employing Sanger sequencing and can complete the verification workflow in less than half the time.

Materials and methods

Samples, library preparation, and sequencing

A primary pancreatic ductal adenocarcinoma and matching normal tissue were obtained from the Australian Pancreatic Cancer Genome Initiative. DNA was extracted using the Allprep DNA/RNA Mini Kit method (Qiagen, Victoria, Australia). A long mate-pair library for each sample was generated according to the Mate-Paired Library Preparation 5500 Series SOLiD Systems kit protocol (Life Technologies, Foster City, CA) (http://tools.lifetechnologies.com/content/sfs/manuals/cms_093442.pdf). Briefly, 5 μg of genomic DNA was sheared into ~2 kb fragments (Covaris S220 System; Life Technologies) and circularized with linker sequences, followed by digestion of the circularized DNA to generate a template used in emulsion PCR. The template was flanked with adaptor sequences, coupled to beads, and clonally amplified before immobilizing it to a solid surface for a 50 bp sequencing run using the SOLiD v4 (Life Technologies).

Analysis of potential somatic rearrangements

Sequence data were mapped to a reference genome based on the Genome Reference Consortium (www.ncbi.nlm.nih.gov/projects/genome/assembly/grc/human/) GRCh37 assembly using the Bioscope v1.2.1 software suite (Applied Biosystems, Foster City, CA). The average physical coverage of the analyzed samples was 268× (tumor) and 211× (normal).

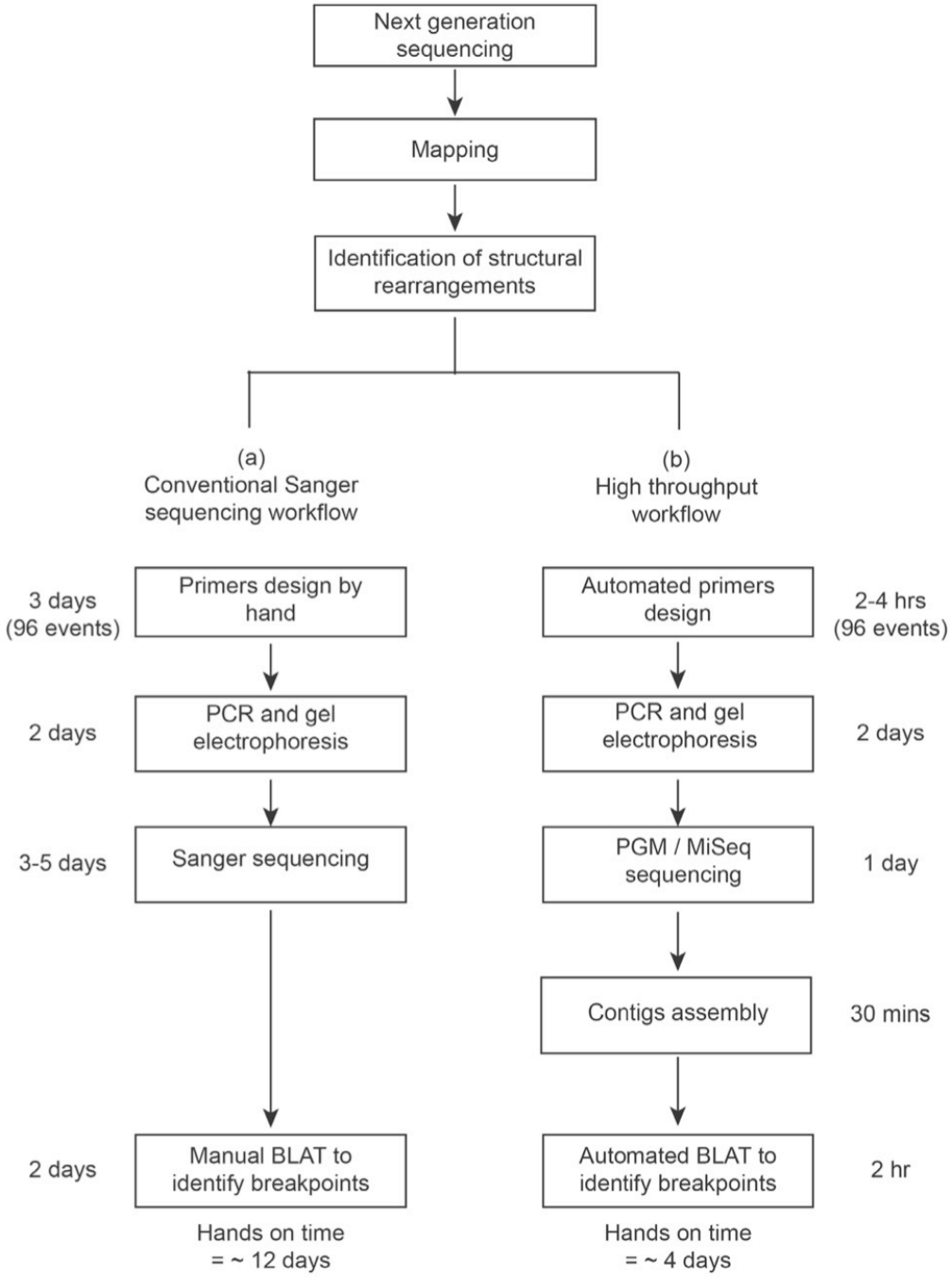


Figure 1. Conventional and high-throughput workflows for the verification of somatic rearrangements and identification of breakpoints. (a) Conventional workflow for verifying somatic rearrangements and their breakpoints using Sanger sequencing. (b) High-throughput verification workflow using either PGM or MiSeq sequencing together with in-house bioinformatics tools to accelerate verification rate. The time to perform each step to verify true somatic rearrangements (assuming 96 events are tested) and determine the sequence of breakpoints is shown. Note that this workflow is applicable for long mate-pair and paired-end whole genome sequencing.

The base coverage was 32× (tumor) and 26× (normal). After mapping, the insert size range for each sequence run was determined to be 630–2780 bp and 670–2900 bp in the tumor and normal tissue, respectively. Discordant read pairs were clustered to identify potential somatic rearrangements using an in-house tool, qSV (Patch et al., manuscript in preparation). A potential rearrangement is supported by at least 10 read pairs. Rearrangements were classified as intra-chromosomal (events within a single chromosome) or inter-chromosomal (events between two chromosome, such as translocations). In cases, where intra-chromosomal rearrangements could be associated with copy number change, together with pair orientation information, this allowed a more specific categorization of the events into deletions, tandem duplications, and inversions. The latter events were not associated with a copy number change and thus were broadly grouped as intra-chromosomal rearrangements. Potential somatic rearrangements were verified using PCR and sequencing (such as benchtop next-generation sequencing and Sanger sequencing).

Automated primer design

The primers were designed using the qAmplicon tool (http://sourceforge.net/p/adamajava/wiki/qAmplicon/) (Lynn et al., manuscript in preparation), which can automatically design primers for each event type (deletions, tandem duplications, intra-chromosomal rearrangements, inversions, and translocations). To do this, qAmplicon extracts a user-defined number of bases with a sequence that spans the predicted breakpoint, it then employs Primer 3 (15) to suggest primer pairs that surround the breakpoint. qAmplicon uses BLAST to match the primers against the reference genome and selects the best primer pairs according to a defined set of criteria, including unique alignment to the reference genome, inability of the primers to form dimers, and an acceptable predefined melting temperature. The input files for qAmplicon include a reference genome and a list of two genomic ranges for each structural rearrangement. Primer pairs for Sanger sequencing were designed to generate a maximum amplicon size of either 1 kb (short amplicon category) or 3 kb (long amplicon category) using benchtop next-generation sequencing.

PCR verification

Each candidate somatic rearrangement was verified by PCR amplification of tumor and matching normal tissue DNA using a 25 µL reaction and was set up in 96 well plates using a robot (Bravo; Agilent Technologies, Victoria, Australia). PCR reactions were performed using the following parameters: 94°C × 2 min, (94°C × 30 s, 60°C × 30 s, 68°C × 1 min) for 40 cycles, 68°C × 15 min. PCR products were visualized by gel electrophoresis (Supplementary Figure S1). A single clear PCR band specific to the tumor at the expected size range was classified as true somatic; PCR bands amplified in both tumor and normal tissue DNA were classified as germline; no band or multiple bands were classified as negative (Supplementary Figure S2).

Sanger sequencing and identification of breakpoints

PCR products of somatic events were purified using AMPure XP (Agencourt, New South Wales, Australia) using a bead:DNA volume ratio of 1.8:1. The products were then quantified using Qubit Flurometer (Invitrogen, Victoria, Australia). Each amplicon was individually prepared for sequencing using forward and reverse primers to increase the chance of identifying the breakpoints. Sequences of verified somatic rearrangements were aligned to the reference genome using BLAT (27) to resolve the breakpoint sequences to base pair level.

Next-generation sequencing and identification of breakpoints

Somatic PCR products ranging up to 3 kb were pooled in equal volumes and purified using AMPure XP bead clean-up at a bead:DNA volume ratio of 1.8:1. Pools of PCR products were quantified using a Qubit Flurometer (Invitrogen). Pooled amplicons were sequenced using the PGM (Life Technologies) and MiSeq (Illumina, San Diego, CA) platforms.

For PGM sequencing, the pool of PCR products was sheared using Ion Shear Plus as per the manufacturer's instructions, then analyzed on an Agilent Bioanalyzer DNA High-Sensitivity LabChip to verify size and purity. Ion libraries were prepared using 50–100 ng of sheared material with the Ion Xpress Plus

Table 1. Summary of designed primers and verification rates for a pancreatic cancer genome using qAmplicon and PCR analysis.

SV	# of potential events	# of long* primers	# of short* primers	# of overlapped primers	# of verified events—long primers (verification rate)	# of verified events—short primers (verification rate)	# of verified events—overall (verification rate)
Deletion	41	41	38	38	17 (39.0%)	19 (50.0%)	19 (46.3%)
Tandem dup	5	5	5	5	1 (20.0%)	1 (20.0%)	1 (20.0%)
Intra-chr	243	226	220	205	147 (65.0%)	166 (75.5%)	183 (75.3%)
Inversion	23	22	21	20	9 (40.9%)	10 (47.6%)	11 (47.8%)
Translocation	46	46	43	43	33 (71.7%)	30 (69.8%)	34 (73.9%)
Total (average rate)	358	340 (95.0%)	327 (91.3%)	311 (86.9%)	207 (60.6%)	226 (69.1%)	248 (69.3%)

* Long primers are suitable for next-generation sequencing and allow flexibility in PCR product size; short primers are suitable for Sanger sequencing restricted to a PCR product size < 1kb.

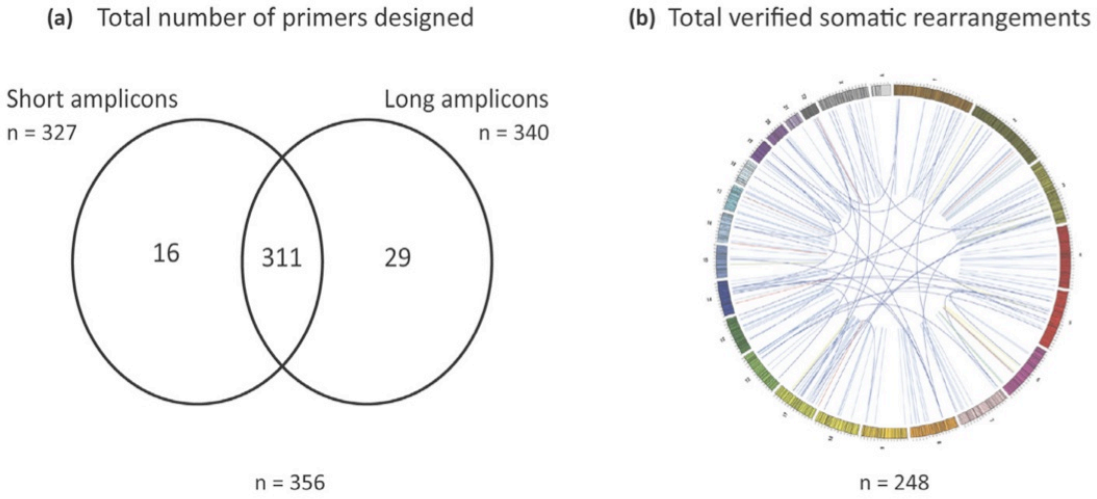


Figure 2. Verification of somatic structural rearrangements from a highly rearranged cancer genome. A total of 356 candidate events had primers designed. (a) A Venn diagram shows the number of primers designed for Sanger sequencing and benchtop next-generation sequencing. (b) A circos image displaying the 248 somatic rearrangements that were verified by Sanger sequencing and/or next-generation sequencing methods. The chromosomes are displayed in the outer ring and the structural rearrangements are shown by the inner connecting lines: green = deletion, red = tandem duplication, orange = inversion, light blue = intra-chromosomal rearrangement, dark blue = translocation.

Fragment Library kit and the AB Library Builder System as per the manufacturer's instructions. Emulsion PCR, emulsion breaking, and enrichment were performed using the Ion OneTouch System to obtain Ion Sphere Particles (ISPs). Enriched ISPs were loaded on an Ion 318 chip on the PGM machine for 200 bp single-read sequencing.

For MiSeq sequencing, the pool of PCR products (1 ng of DNA) was prepared using the Nextera XT kit. The DNA was simultaneously fragmented and tagged with adaptors using a tagmentation enzymatic reaction, followed by PCR amplification. Sequencing was performed on the MiSeq system with paired-end 2 × 150 bp reads.

After sequencing, the short reads were assembled to create consensus contigs using an in-house de novo assembly tool (qNovo4.pl; Supplementary File 1), which uses an overlap consensus method. However, any other assembly tool, such as EBARDenovo (16), Oases (17), or Trinity (18), could be used in this workflow. The first step of the qNovo4 assembly trimmed the reads based on a user-defined Phred score and ambiguous base (N) content. The reads were then filtered by nucleotide complexity using user-defined Shannon's entropy for mono-, di-, and tri-nucleotide content. Reads that did not pass the filters or reads that were shorter than the user-defined minimum after trimming were not used in the assembly. New contigs generated by sequence artifacts were avoided by first collapsing all reads that were within Hamming distance based on the user-defined mismatch rate. The construction of each contig began with a randomly selected read. All reads with a minimum overlap length and less than the maximum Hamming distance were aligned to construct a consensus sequence. The number of reads (based on user defined threshold) used to construct the consensus sequence justified the contig extension. Extension occurs from both the 5′ and 3′ ends of the contig. The built contig was used as a seed for another round of extension, until there were no reads justifying an extension. Reads used to generate the contig were removed from the memory and were not available to form other contigs. qNovo4 reiterated these steps for every read in the memory. The assembled contigs generated in this workflow used the following parameters: minimum length of the tags = 30 bp, mismatch rate = 3, and length of seed = 20 bp. Resultant contigs were automatically aligned to the human reference genome (GRCh37/hg19) using a local installation BLAT. Aligned results were searched for matching blocks with the correct orientation in the predicted genomic region. The identification of breakpoints was based on the highest quality BLAT match that was selected to maximize the length of the contig and the score of matched bases.

Results and discussion

The aim of this study was to develop a workflow of standard techniques to increase the throughput process for verifying structural rearrangement events and identifying breakpoint locations (Figure 1). To evaluate the workflow, we sequenced DNA from a tumor and matched normal tissue, identifying a total of 358 potential somatic rearrangements in this cancer genome using a threshold of at least 10 read pairs supporting an event. Of these, 41 (11.5%) were deletions, 5 (1.4%) tandem duplications, 243 (67.9%) intra-chromosomal rearrangements, 23 (6.4%) inversions, and 46 (12.9%) translocations (Table 1).

We were able to design primers to verify 327/358 (91.3%) of the candidate

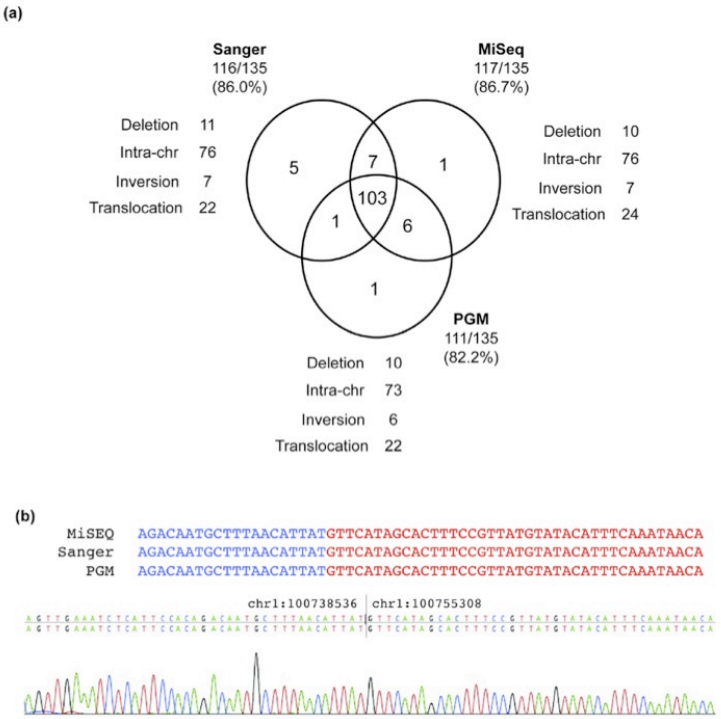


Figure 3. The ability of different sequencing approaches to resolve the sequences of breakpoints. A total of 135 events were tested using Sanger and next-generation sequencing platforms (PGM and MiSeq). In all cases, the short amplicon (~1 kb PCR product size) were used. (a) Number of identified breakpoints across the three sequencing platforms. Venn diagram shows the respective number of identified breakpoints from the individual platform and the overlap breakpoints. (b) Representative breakpoint sequence from an intrachromosomal rearrangement is confirmed by PCR and sequencing showing 100% identity when aligning sequences from the three platforms. The identified breakpoint is chr1:100738536-100755308 with a resolution of 16,773 bp.

rearrangement events by Sanger sequencing. However, due to the low complexity of DNA sequences near the breakpoints (19), qAmplicon was not able to generate primer pairs for all predicted events. Similar problems were also encountered when primers pairs were designed manually using Primer 3 without the filtering performed by qAmplicon. In contrast, the advantage of amplicon sequencing with a next-generation sequencer (PGM and MiSeq), is that unlike Sanger sequencing where primer design is limited to an amplicon size of <1 kb (short amplicons), there is no limitation on amplicon size. This is because the amplicons are fragmented, sequenced, and assembled prior to aligning to the reference genome. Therefore, by allowing PCR products of up to 3 kb, an extra set of 29 primers could be designed (classified as long amplicons), allowing more events to be tested (Figure 2a, Table 1). Altogether, qAmplicon generated a total of 356 (99.4%) primers from 358 potential somatic rearrangements (Supplementary Table S1). There were 311/356 designed primers shared between the short amplicon and long amplicon categories (Figure 2a).

The verification rate of the Sanger primers (short amplicons) was assessed by PCR and gel electrophoresis. Of the 327 primers, a total of 226 (69.1%) events were confirmed as somatic, 13 were confirmed as germline (4.0%), and 88 (27.0%) events were negative. However, an extra 22 events were confirmed as somatic using next-generation sequencing primers (long amplicons). Collectively, a total of 248 (69.0%) somatic events were verified, which included 19 deletions, 1 tandem duplication, 183 intra-chromosomal rearrangements, 11 inversions, and 34 translocations (Figure 2b, Table 1). Interestingly, the PCR results revealed that 62 of the events tested with both long and short amplicon primers yielded different results (Supplementary Figure S3). This suggests that primer design is crucial, and a negative result may not necessarily indicate that an event is false; thus, some PCRs may require individual optimization in order to verify an event.

The sequences surrounding the breakpoints of somatic rearrangements can reveal insights into their formation by identifying potential mechanisms of DNA damage repair (20). Therefore, we compared the ability of the conventional Sanger and high-throughput next-generation sequencing workflow to resolve the sequences of breakpoints using a subset of 135 events. Due to the limitations of Sanger sequencing, only the short amplicons were used to generate the amplicons for sequencing with each platform. The assembly generated a total of 5906 contigs from PGM sequencing data (range 101–1289 bp) and 7866 contigs from MiSeq sequencing data (range 101–6507 bp). Of these, a total of 124 (91.9%) breakpoints were identified by at least 1 sequencing method. Sanger sequencing identified 116 (86.0%), MiSeq sequencing identified 117 (86.7%), and PGM sequencing identified 111 (82.2%) (Figure 3a). Overall, 103 (76.3%) breakpoints were identified by all 3 methods.

To determine the accuracy of the benchtop next-generation sequencing platforms (PGM and MiSeq) to resolve the breakpoints, we compared the breakpoint sequences of 103 events that were identified by both benchtop next-generation sequencers to Sanger sequencing (Figure 3b). The comparison showed that MiSeq sequencing accurately identified 100 of 103 breakpoints (97.1%), while PGM sequencing identified 95 breakpoints (92.2%). The reason the exact locations for some breakpoints were not identified is likely due to the gapped alignment of the contigs with homopolymers (Supplementary Figure S4). In this context, the

assembled contigs obtained by MiSeq sequencing were slightly more accurate than those obtaing by PGM sequencing, possibly due to the paired-end information and fewer sequence errors around homopolymeric sequences. The observed performance of the two benchtop next-generation sequencing platforms is consistent with previous studies (21,22). However, MiSeq and PGM identified eight additional breakpoints that were not identified by Sanger sequencing (Figure 3a). These results showed that, with respect to the number of identified breakpoint sequences, the high-throughput workflow using MiSeq and PGM sequencing methods showed high accuracy and was comparable to conventional Sanger sequencing.

In conclusion, we have described a high-throughput workflow for verifying somatic structural rearrangements in cancer genomes. The advantages of this high-throughput workflow include (*i*) the utilization of benchtop next-generation sequencing (Ion Torrent PGM and Illumina MiSeq) to replace conventional Sanger sequencing, which mitigates the concern of PCR product size, allowing primers to be designed for more events; (*ii*) the integration of bioinformatics tools and next-generation sequencing based methods greatly increases the speed and volume of the verification process; and (*iii*) the accuracy of the next-generation sequencing methods is comparable to that of the conventional Sanger sequencing method. Although the results presented here were based on the detection of structural rearrangements from SOLiD sequencing data, this workflow is also applicable to Illumina HiSeq data as the qSV tool has the ability to identity structural rearrangements in both platforms (Supplementary Table S2).

We expect that this workflow will enable routine verification of somatic rearrangements and breakpoints, enhancing large-scale screening projects. This workflow can be useful in characterizing rearrangement breakpoints that may lead to disrupted repair pathways and could help prioritize treatment options for patients in future personalized medicine applications (23). The rapid verification of somatic events can also allow the identification of tumor-specific breakpoints as potential candidate biomarkers in cancer patients to assess disease progression in real time (24–26).

Author Contributions

All authors contributed to the work: K.Q., K.N., and N.W. wrote the manuscript; K.Q., K.N., and N.W. conceived the experiments; A.M.P., J.L.F., F.N., N.C., and K.K. contributed to the qAmplicon software; K.Q., D.M., M.F., A.C., T.B., S.M., I.H., S.I., C.N., E.N., S.W., and A.S. performed the laboratory experiments and next-generation sequencing; M.A., O.H., C.L., D.T., S.W., Q.X., and J.V.P. performed variant calling and data management of the next-generation sequence data; the APGI and A.V.B. contributed tumour samples; K.N., P.W., N.W., and S.M.G. oversaw the work.

Acknowledgments

We would like to thank D. Gwynne for central coordination at the Queensland Centre for Medical Genomics. We wish to thank all of the APGI members as well as patients who contribute to APGI. This research has been supported by the National Health and Medical Research Council of Australia (NHMRC: 631701, 535903, 427601, APP1047334); Australian Government: Department of Innovation, Industry, Science and Research (DIISR); Australian Cancer Research Foundation (ACRF); Queensland Government (NI-RAP); University of Queensland; Cancer Council NSW: (SRP06-01); Cancer Institute NSW: (10/ECF/2-26; 06/ECF/1-24; 09/CDF/2-40; 07/CDF/1-03; 10/CRF/1-01, 08/RSA/1-15, 07/CDF/1-28, 10/CDF/2-26,10/FRL/2-03, 06/RSA/1-05, 09/RIG/1-02, 10/TPG/1-04, 11/REG/1-10, 11/CDF/3-26); Garvan Institute of Medical Research; Avner Nahmani Pancreatic Cancer Research Foundation; R.T. Hall Trust; Petre Foundation; Philip Hemstritch Foundation; Gastroenterological Society of Australia (GESA); American Association for Cancer Research (AACR) Landon Foundation INNOVATOR Award; Royal Australasian College of Surgeons (RACS); Royal Australasian College of Physicians (RACP); Royal College of Pathologists of Australasia (RCPA). S.G. is a recipient of a NHMRC Principal Research Fellowship.

Competing Interests

The authors declare no completing interests.

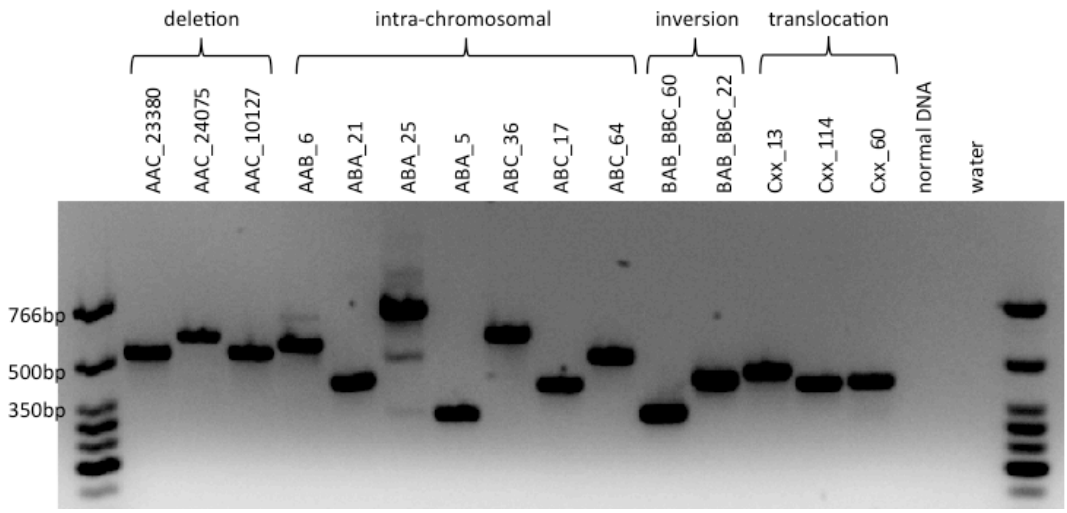
References

1. **Stratton, M.R., P.J. Campbell, and P.A. Futreal.** 2009. The cancer genome. Nature *458*:719-724.
2. **Negrini, S., V.G. Gorgoulis, and T.D. Halazonetis.** 2010. Genomic instability–an evolving hallmark of cancer. Nat. Rev. Mol. Cell Biol. *11*:220-228.
3. **Leary, R.J., J.C. Lin, J. Cummins, S. Boca, L.D. Wood, D.W. Parsons, and S. Jones. T. Sjöblom, et al.** 2008. Integrated analysis of homozygous deletions, focal amplifications, and sequence alterations in breast and colorectal cancers. Proc Natl Acad Sci U S A. *105*:16224-16229.
4. **Nowell, P.C. and D.A. Hungerford.** 1960. Minute chromosome in human chronic granulocytic leukemia. Science *142*:1497.
5. **Sadikovic, B., K. Al-Romaih, J.A. Squire, and M. Zielenska.** 2008. Cause and consequences of genetic and epigenetic alterations in human cancer. Curr. Genomics *9*:394-408.
6. **Xing, E.P., G.Y. Yang, L.D. Wang, S.T. Shi, and C.S. Yang.** 1999. Loss of heterozygosity of the Rb gene correlates with pRb protein expression and associates with p53 alteration in human esophageal cancer. Clin Cancer Res. *5*:1231-1240.
7. **Chen, K., J.W. Wallis, M.D. McLellan, D.E. Larson, J.M. Kalicki, C.S. Pohl, S.D. McGrath, M.C. Wendl, et al.** 2009. BreakDancer: an algorithm for high-resolution mapping of genomic structural variation. Nat. Methods *6*:677-681.
8. 8. **Ye, K., M.H. Schulz, Q. Long, R. Apweiler, and Z. Ning.** 2009. Pindel: a pattern growth approach to detect break points of large deletions and medium sized insertions from paired-end short reads. Bioinformatics *25*:2865-2871.
9. **Wang, J., C.G. Mullighan, J. Easton, S. Roberts, S.L. Heatley, J. Ma, M.C. Rusch, K. Chen, et al.** 2011. CREST maps somatic structural variation in cancer genomes with base-pair resolution. Nat. Methods *8*:652-654.
10. **The International Cancer Genome Consortium.** 2010. International network of cancer genome projects. Nature *464*:993-998.
11. **Cancer Genome Atlas Research Network.** 2011. Integrated genomic analyses of ovarian carcinoma. Nature *474*:609-615.
12. **Campbell, P.J., P.J. Stephens, E.D. Pleasance, S. O'Meara, H. Li, T. Santarius, L.A. Stebbings, C. Leroy, et al.** 2008. Identification of somatically acquired rearrangements in cancer using genome-wide massively parallel paired-end sequencing. Nat. Genet. *40*:722-729.
13. **Hillmer, A.M., F. Yao, K. Inaki, W.H. Lee, P.N. Ariyaratne, A.S.M. Teo, X.Y. Woo, Z. Zhang, et al.** 2011. Comprehensive long-span paired-end-tag mapping reveals characteristic patterns of structural variations in epithelial cancer genomes. Genome Res. *21*:665-675.
14. **Stephens, P.J., D.J. McBride, M.-L. Lin, I. Varela, E.D. Pleasance, J.T. Simpson, L.A. Stebbings, C. Leroy, et al.** 2009. Complex landscapes of somatic rearrangement in human breast cancer genomes. Nature *462*:1005-1010.
15. **Rozen, S. and H. Skaletsky.** 2000. Primer3 on the WWW for general users and for biologist programmers. Methods Mol. Biol. *132*:365-386.
16. **Chu, H.T., W.W. Hsiao, J.C. Chen, T.J. Yeh, M.H. Tsai, H. Lin, Y.W. Liu, S.A. Lee, et al.** 2013. EBARDenovo: highly accurate de novo assembly of RNA-Seq with efficient chimera-detection. Bioinformatics *29*:1004-1010.
17. **Schulz, M.H., D.R. Zerbino, M. Vingron, and E. Birney.** 2012. Oases: robust de novo RNA-seq assembly across the dynamic range of expression levels. Bioinformatics *28*:1086-1092.
18. **Grabherr, M.G., B.J. Haas, M. Yassour, J.Z. Levin, D.A. Thompson, I. Amit, X. Adiconis, L. Fan, et al.** 2011. Full-length transcriptome assembly from RNA-Seq data without a reference genome. Nat. Biotechnol. *29*:644-652.
19. **Wei, Y., M. Sun, G. Nilsson, T. Dwight, Y. Xie, J. Wang, Y. Hou, O. Larsson, et al.** 2003. Characteristic sequence motifs located at the genomic breakpoints of the translocation t(X;18) in synovial sarcomas. Oncogene *22*:2215-2222.
20. **Onishi-Seebacher, M. and J.O. Korbel.** 2011. Challenges in studying genomic structural variant formation mechanisms: the short-read dilemma and beyond. Bioessays. *33*:840-850.
21. **Loman, N.J., R.V. Misra, T.J. Dallman, C. Constantinidou, S.E. Gharbia, J. Wain, and M.J. Pallen.** 2012. Performance comparison of benchtop high-throughput sequencing platforms. Nat. Biotechnol. *30*:434-439.
22. **Quail, M.A., M. Smith, P. Coupland, T.D. Otto, S.R. Harris, T.R. Connor, A. Bertoni, H.P. Swerdlow, and Y. Gu.** 2012. A tale of three next generation sequencing platforms: comparison of Ion Torrent, Pacific Biosciences and Illumina MiSeq sequencers. BMC Genomics *13*:341.
23. **Stephens, P.J., C.D. Greenman, B.Y. Fu, F.T. Yang, G.R. Bignell, L.J. Mudie, E.D. Pleasance, K.W. Lau, et al.** 2011. Massive Genomic Rearrangement Acquired in a Single Catastrophic Event during Cancer Development. Cell. *144*:27-40.
24. **Diehl, F., K. Schmidt, M.A. Choti, K. Romans, S. Goodman, M. Li, K. Thornton, N. Agrawal, et al.** 2008. Circulating mutant DNA to assess tumor dynamics. Nat. Med. *14*:985-990.
25. **Leary, R.J., I. Kinde, F. Diehl, K. Schmidt, C. Clouser, C. Duncan, A. Antipova, C. Lee, et al.** 2010. Development of Personalized Tumor Biomarkers Using Massively Parallel Sequencing. Sci. Transl. Med. *2*:20ra14.
26. **McBride, D.J., A.K. Orpana, C. Sotiriou, H. Joensuu, P.J. Stephens, L.J. Mudie, E. Hamalainen, L.A. Stebbings, et al.** 2010. Use of cancer-specific genomic rearrangements to quantify disease burden in plasma from patients with solid tumors. Genes Chromosomes Cancer *49*:1062-1069.
27. **Kent, W.J.** 2002. BLAT--the BLAST-like alignment tool. Genome Res. *12*:656-664

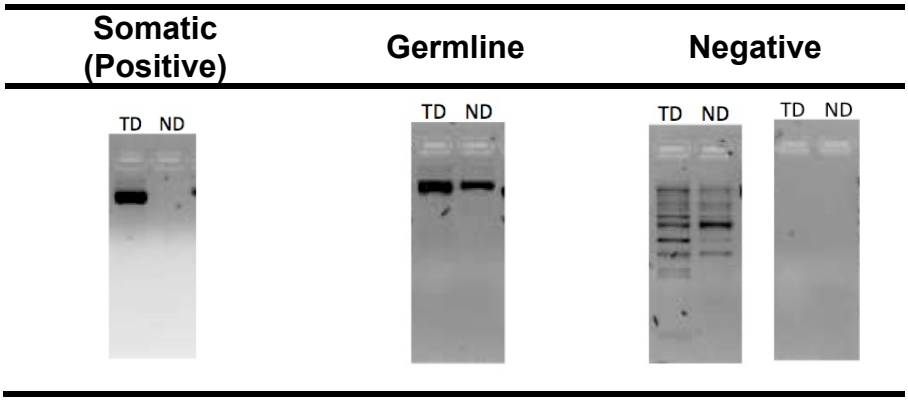
Received 22 April 2014; accepted 12 June 2014.

Address correspondence to Nic Waddell or Sean M. Grimmond, Queensland Centre for Medical Genomics, Institute for Molecular Bioscience, University of Queensland, St Lucia, Brisbane, QLD, Australia. E-mail: n.waddell@imb.uq.edu.au; s.grimmond@imb.uq.edu.au or Sean.Grimmond@glasgow.ac.uk

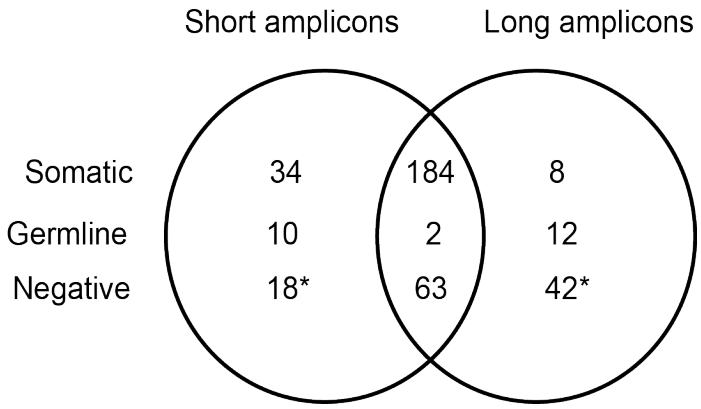
2.3 Supplementary material



Supplementary Figure 2-1 PCR verification of candidate somatic rearrangements using short amplicon primers. This is an example of a gel image of four different event types.



Supplementary Figure 2-2 Classification of PCR verification of candidate somatic rearrangements. The 3 categories obtained after gel electrophoresis of the PCR product spanning the breakpoints of rearrangements in 2% agarose gel are classified as: Somatic (positive) - PCR reactions yield a single, clear PCR band in tumour DNA (TD) with no matching band in the normal DNA (ND); Germline - both tumour and normal yield PCR bands of the expected size; Negative – PCR did not yield any band or PCR yields multiple bands; PCR reaction are inconclusive as primers might not specific to the events or the failure of PCR reactions.



Supplementary Figure 2-3 Breakdown of verification results for 311 candidate somatic events tested by both short and long amplicons. In total 249 events showed the same verification results when tested by short or long amplicons. The remaining 62 events tested both long and short amplicon primers yielded different results (short amplicons = 34+10+18; long amplicons = 8+12+42).

*Note: A negative result may not indicate that the event was not verified. The 18 events called negative in short amplicons category, 8 were somatic and 10 were germline in the other category. Similar for the 42 negative events in the long amplicons category, 32 were somatic and 10 were germline in short amplicons category.

(a) Assembled contig sequences

Sanger

```
TGAATCATGC AACACTTGGG ACCTAACTGG CCGATGCTGG ACCTAACCTG  50
CAAATGCTAG AAAACTCACC ACAACAGAGC AAGCAAAACT CCTTGCAAAG  100
TTTTTGCAGG TCTCTCAGTT TTTGTTTTGT TTTTGTTTTT GTTTTGAGAT  150
GGAGTCTTGC TCTGTCGCCC AGGCCAAAGT GCAGAGGTGC GATCTCGGCT  200
CACTGCAATC TCTGCCTCCC AGGTTCGAGC AATTCTCCTG CCTCAGCCTC  250
CCGAGTAGCT GGGATTACAG GCACGAATC
```

PGM

```
AACCCAGGAC CAACCTCCAA TGCACAGGCT CTTGACTGCA GGGCAAAGGA  50
GGTTGACAGT TTTAGTCCCA ATGAATCATG CAACACTTGG GACCTAACTG  100
GCCGATGCTG GACCTAACCT GCAAATGCTA GAAAACTCAC CACAACAGAG  150
CAAGCAAAAC TCCTTGCAAA GTTTTGCAGG TCTCTCAGTT TTGTTTTGTT  200
TTGTTTTGTT TTGAGATGGA GTCTTGCTCT GTCGCCCAGG CCAAAGTGCA  250
GAGGTGCGAT CTCGGCTCAC TGCAATCTCT GCCTCCCAGG TTCGAGCAAT  300
TCTCCTGCCT CAGCCTCCCG AGTAGCTGGG ATTACAGGCA CGAATCACCA  350
CACCTGGCTT A
```

MiSeq

```
ggttCCCAGG ACCAACCTCC AATGCACAGG CTCTTGACTG CAGGGCAAAG  50
GAGGTTGACA GTTTTAGTCC CAATGAATCA TGCAACACTT GGGACCTAAC  100
TGGCCGATGC TGGACCTAAC CTGCAAATGC TAGAAAACTC ACCACAACAG  150
AGCAAGCAAA ACTCCTTGCA AAGTTTTTGC AGGTCTCTCA GTTTTTGTTT  200
TGTTTTTGCT TTTGTTTTTGA GATGGAGTCT TGCTCTGTCG CCCAGGCCAA  250
AGTGCAGAGG TGCGATCTCG GCTCACTGCA ATCTCTGCCT CCCAGGTTCG  300
AGCAATTCTC CTGCCTCAGC CTCCCGAGTA GCTGGGATTA CAGGCACGCA  350
```

(b) Alignment of MiSeq and Sanger sequence

```
MiSeq  ...ACTCACCACAACAGAGCAAGCAAAACTCCTTGCAAAGTTTTT...
Sanger ...ACTCACCACAACAGAGCAAGCAAAACTCCTTGCAAAGTTTTT...
```

Alignment of PGM and Sanger sequence

```
PGM    ...ACTCACCACAACAGAGCAAGCAAAACTCCTTGCAAAGTTTT-GCAGGTCTCTCA
Sanger ...ACTCACCACAACAGAGCAAGCAAAACTCCTTGCAAAGTTTTTGCAGGTCTCTCA

PGM    GTTTT-GTTTTGTTTT-GTTTT-GTTTTGAGATGGAGTCTTGCTCTGTCGCCCAGGCCAA
Sanger GTTTTTGTTTTGTTTTTGTTTTTGTTTTGAGATGGAGTCTTGCTCTGTCGCCCAGGCCAA

PGM    AGTGCAGAGGTGCGATCTCGGCTCACTGC
Sanger AGTGCAGAGGTGCGATCTCGGCTCACTGC
```

Supplementary Figure 2-4 An illustration of the gapped alignment of the sequencing reads taken from an intra-chromosomal rearrangement. (a) Assembled contigs sequences obtained from Sanger, MiSeq and PGM sequencing. The nucleotides highlighted in green and red denote the breakpoints of the rearranged DNA fragments. Nucleotides in black indicate non-templated sequences insertion. The underlined nucleotides mark the stretches of Ts called by the three sequencing methods. (b) The alignment of assembled contigs was compared to Sanger sequences. In this example, the breakpoint identified by MiSeq and Sanger sequencing was at chr17:70371942-70425723 with T nucleotide insertion. While the breakpoint identified by PGM was at chr17:70371950-70425723 with TTGCAAAG nucleotides insertion.

3 Chapter 3

Characteristics of somatic breakpoints may indicate repair mechanisms that were active or absent during the generation of genomic rearrangements

3.1 Introduction

Structural rearrangements are large genetic events in the genome, which may have significant impact on genotypic and phenotypic modifications of tumour cells. The recent advancement of sequencing technologies allows us to identify a large range of somatic rearrangements to base pair resolution.

In the previous chapter, I established a high throughput workflow to rapidly verify somatic rearrangements and confirmed that the qSV tool was able to detect the exact location of breakpoints and accurately identify the DNA sequence at each breakpoint junction. In this current chapter, I have conducted a detailed analysis of the rearrangement breakpoints obtained from whole genome sequencing data of 120 primary pancreatic primary tumours. This analysis is able to inform the processes which may have occurred during the formation of the breakpoints. Such process may include the defective DNA repair mechanisms and may represent candidate targets for therapy.

Double-stranded DNA repair

In general, there are two main DNA repair mechanisms that cope with double-stranded DNA breaks (DBS): homology-dependent and homology-independent mechanisms. Defects in these mechanisms are known to be involved in the formation of somatic rearrangements.

The homology-dependent mechanism (also known as homologous recombination, HR) is a DNA repair process that uses extensive sequence identity between DNA fragments. This identity generally extends 100 – 200 bp (Chen, 2001) and allows for the accurate repair of DSB during cell cycle. HR is regulated by a number of genes and the BRCA1 and BRCA 2 genes are recognized as key players in the HR pathway (Ishioka et al., 1997).

Furthermore, the *BRCA1/2* genes are also associated with genomic instability in cancer, as the inactivation of *BRCA1/2* leads to an increase in number of large structural rearrangements, which are at least 10 Mb in size (Popova et al., 2012).

Homology-independent mechanism, also termed non-homologous end joining (NHEJ), is an alternate DNA repair mechanism, which re-joins the broken ends of DNA either without the need of any sequence homology (0 bp; 'accurate' NHEJ) or with microhomology (in the range of 1 – 25 bp) (Chen, 2001). There are two sub-pathways that utilise microhomology, they are 'error-prone' NHEJ and microhomology-mediated end joining (MMEJ).

'Error prone' NHEJ uses short microhomology (1 to 5 bp) present at single stranded overhang of the DNA break to facilitate repair. Whilst in MMEJ, the end joining event reveals microhomology of 6 to 25 bp to guide the restoration of the break (Chen, 2001). Recent cancer studies have shown that the use of microhomology to facilitate repair of DNA break is highly prevalent in human cancers (Campbell et al., 2008; Hillmer et al., 2011; Ng et al., 2012; Stephens et al., 2009).

Mobile DNA elements at rearrangement breakpoints

In addition to the defective DNA repair pathways, other mechanisms may result in structural rearrangements. Mobile elements are abundant in human genome and account for approximately 20% of germline structural rearrangements among individuals (Kidd et al., 2010; Xing et al., 2009). Mobile elements can replicate and move within the genome. These elements can replicate through RNA intermediates or "cut-and-paste" mechanism (Finnegan, 1989). Mobile elements can be broadly divided into 2 classes: retrotransposons and DNA transposons. Retrotransposons produce RNA transcripts, which reverse transcribe into DNA sequences and then insert into a target site. In contrast, DNA transposons rely on human transposase enzymes to cut on the target site and then move from one to another region in genome.

In tumours, it has been shown that there is a high density of mobile elements at some tumour specific breakpoints and these may facilitate cancer-inducing structural rearrangements such as deletions, duplications or translocations (Mauillon et al., 1996; Montagna et al., 1999; Rowe et al., 1995; Strout et al., 1998). Given their nature, the

activity of transposable elements is mostly suppressed by epigenetic mechanisms at transcriptional and post-transcriptional levels (Slotkin and Martienssen, 2007; Yang and Kazazian, 2006). However, in the context of disease, particularly in cancer (such as colorectal, prostate, breast and ovarian) (Debniak et al., 2001; Depil et al., 2002; Lee et al., 2012; Montagna et al., 1999; Solyom et al., 2012), it has been shown that the activity of transposable elements is frequently elevated. This suggests that the suppression mechanisms have been disrupted allowing the transposable elements to regain the mobility and may generate new genotypic and phenotypic modifications in the genome. Presently, it is uncertain whether these mobile elements are initiators of tumours development or are activated later.

Structural rearrangements in pancreatic cancer

Recent work from our laboratory demonstrated that the pattern of somatic rearrangements is able to classify tumours into different genome subtypes with potential clinical relevance (Waddell et al., 2015). The pancreatic cancer tumours were classified as follows:

- *Stable* – tumours with low number of somatic rearrangements distributed across the genome (<50 events)
- *Scattered* – tumours contain a modest number events (50-200 events)
- *Focal* – tumours whereby >50% events occur on a single chromosome
- *Unstable* – tumours harbour more than 200 events across the entire genome

The unstable subtype was associated with somatic *BRCA2* or germline *BRCA2* and *PALB2* mutations. Furthermore, the unstable phenotype was suggested as a candidate biomarker of therapeutic responsiveness to certain platinum-based chemotherapies which generate DNA breaks, as it was hypothesised that these tumours have a defective HR pathway. Tumours with a defective HR pathway have been shown to be sensitive to platinum-based therapies or PARP inhibitors (Farmer et al., 2005), thus the identification of which tumours harbour a defective HR pathway is clinically important. However, the unstable subtype or *BRCA* mutation alone was not able to predict a defective HR pathway with 100% accuracy as some tumours which were genomically unstable, did not have a BRCA signature or *BRCA* mutation which suggests other mechanisms are driving the instability in these cases. While some tumours with no mutation in BRCA genes or other members of HR pathway contained a high BRCA signature or defective HR. Furthermore, in some instances, there may be a *BRCA* mutation which is of unknown function or clinical

significance, therefore it is unclear whether the tumours are likely to have a defective HR pathway and thus respond to therapy. As the breakpoint characteristics may reflect the types of DNA repair which are occurring, in this chapter, I tested whether the characteristics at breakpoints may indicate which tumour have a defective *BRCA* pathway in pancreatic cancer. The analysis of breakpoint characteristics also included two other cancers types (i.e. oesophageal and ovarian cancers) to further understand the pattern of somatic breakpoints characteristics in tumours with defective HR.

Apart from studying the breakpoints, I analysed the DNA sequences surrounding the rearrangements to identify DNA sequence enrichment (i.e. motif) associated with the breakpoints. Many of the somatic rearrangements breakpoints displayed strong enrichment of A+Ts suggesting that retrotransposition activity and/or common fragile sites may implicate in pancreatic tumorigenesis.

3.2 Material and Methods

This chapter utilised whole genome sequence data, which was generated as part of the pancreatic and ovarian ICGC projects. The data was sequenced, mapped and qSV tool was run to identify all structural rearrangements present in each tumour. This work was performed by members of the ICGC pancreatic team. I then used the output of qSV to carry out the analysis. The breakpoint sequences were obtained by qSV from the contigs assembled from abnormally mapped soft clipped and unmapped reads around potential breakpoints. I classified each breakpoint in to 5 different phenotypes (blunt end, microhomology, non-templated insertion, genomic shard and free DNA with repeat annotation) (Figure 3-1). To understand more about the mechanisms of DNA damage repair, the proportion of each breakpoint characteristic and the actual sequence surrounding the breakpoints was analysed across the patient cohort. An outline of the experimental workflow is shown in Figure 3-2.

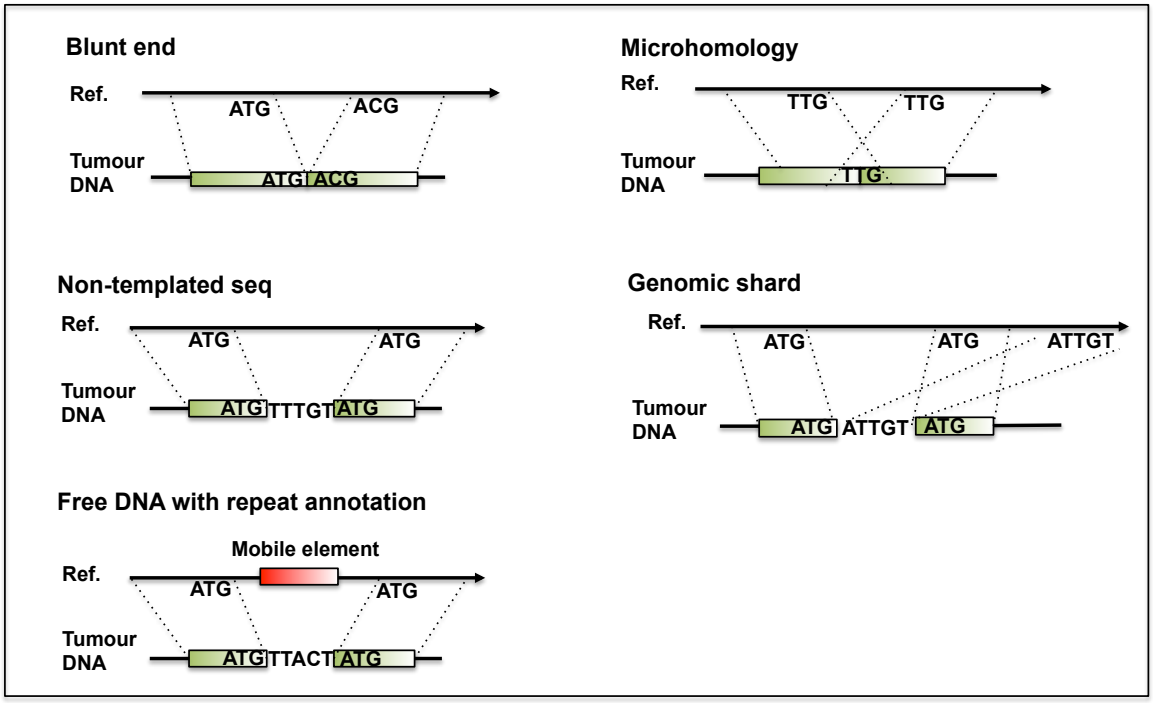


Figure 3-1 Characteristics of rearrangements breakpoints. The breakpoints are characterised into 5 different types: blunt end - a clean break whereby the broken chromosome ends are re-joined accurately, microhomology - the presence of terminal short overlapping bases between the two repaired chromosome ends, non-templated sequence insertion - inserted sequences of free DNA found at the junction of breakpoints, genomic shard - inserted free DNA that mapped uniquely to the reference genome near the vicinity of the breakage or elsewhere in the genome and free DNA with repeat annotation - inserted free DNA that overlapped the sequence of mobile elements.

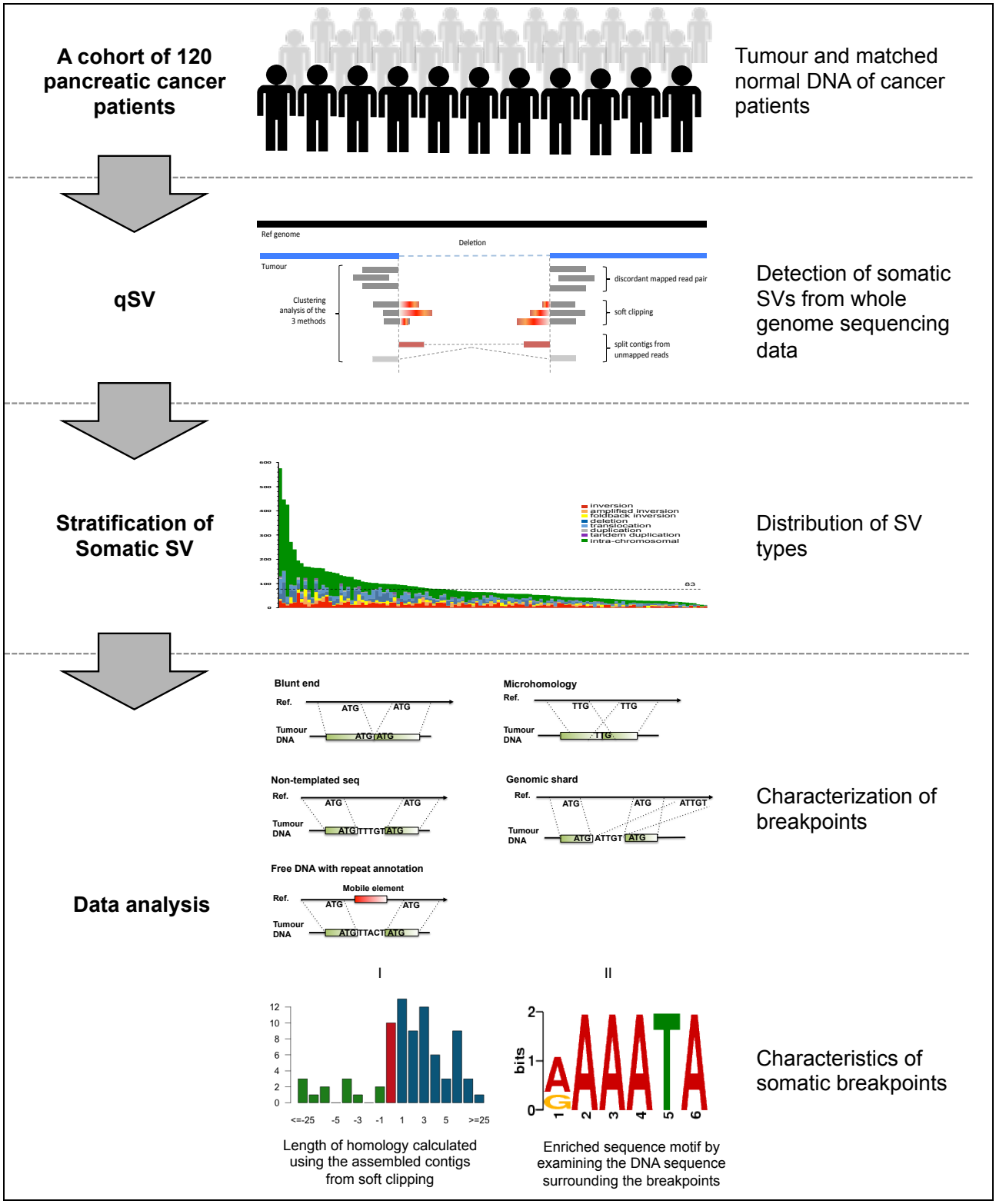


Figure 3-2 Overview framework to analysis the characteristics of somatic rearrangements breakpoints. Whole genome sequencing data of 120 primary tumours and their matched normal DNA were analysed by the ICGC team using the qSV tool to detect potential somatic rearrangements. The somatic rearrangements were categorized into the different event types including inversions, amplified inversions, foldback inversions,

deletions, translocations, duplications, tandem duplications, and intra-chromosomal rearrangements. Downstream analysis of assembled contigs at the breakpoints allows the characterization of the breakpoints junctions into 5 phenotypes: blunt end, microhomology, non-templated insertion, genomic shard and free DNA with repeat annotation. The length of homology at the breakpoints was used to identify potential active repair mechanisms in cancers. The DNA sequences surrounding the breakpoint junctions from the reference genome were used to identify DNA enriched sequence motifs.

3.2.1 Sample, library preparation and sequencing

Pancreatic ductal adenocarcinoma and matching normal tissue were obtained from the Australian Pancreatic Cancer Genome Initiative (APGI). DNA was extracted using the Allprep DNA and RNA Qiagen Allprep® Kit in accordance with the manufacturer's instructions (Qiagen). Paired end library for each sample was constructed according to TruSeq® DNA LT Sample Prep Kit v2. In brief, 1 µg genomic DNA was fragmented into approximately 300 bp using the Covaris™ S2 sonicator. The fragmented DNA were prepared into libraries using the standard illumina© library preparation workflow. The fragments DNA were end-repaired, added a 3'-A overhang and ligated with indexed adapter. The adapter ligated fragments were size selected with two rounds of SPRI beads purification (AxyPrep™Mag PCR Clean-up) using a final bead to DNA volume ratio of 0.60:1 followed by 0.70:1. An average size of 500 bp molecules were selected and then amplified with a total of 8 cycles of PCR to generate constructs compatible for 2 x 101 bp HiSeq sequencing. ***Note: Sequencing was performed by QCMG sequencing team.***

3.2.2 Data pre-processing

Sequence data was aligned to a reference human genome based on the Genome Reference Consortium GRCh37 assembly using BWA (Li and Durbin, 2009). The alignments were converted to the sequence alignment/mapping (SAM) format and then compressed into binary file (BAM) using Samtools (Li et al., 2009). Sorted BAM files were merged into a single BAM for each tumour and normal sample. PCR duplicates were marked using PiCard MarkDuplicates http://picard.sourceforge.net. The alignment summary statistics such as coverage estimation for the merged BAM files were carried out by in-house tools. These tools are available for download at http://sourceforge.net/projects/adamajava. ***Note: Data pre-processing were performed by QCMG bioinformatics team.***

3.2.3 Identification of potential somatic rearrangements and breakpoints

All potential somatic rearrangements were identified using qSV (http://sourceforge.net/projects/adamajava). qSV provides three lines of evidence to call potential somatic structural rearrangements – clusters of discordant mapped read pairs, soft clipping and a split contig alignment generated from de novo assembly of unmapped and aberrantly mapped reads. All high confidence events were used in downstream analysis and are defined by the following criteria:

- The presence of multiple lines of evidence (discordant pairs, soft clipping on both ends and split reads)
- The presence of two lines of evidence: discordant pairs on both breakpoints and soft clipping; discordant pairs on both breakpoints and split reads; or soft clipping on both ends and split reads
- Clustering of 10 or more reads with the same characteristic was used as a cut-off to support a potential event.

3.2.4 Verification of potential somatic rearrangements and breakpoints

Verification of events by SOLiD next generation sequencing was performed by QCMG sequencing team. Essentially, 823 of 1,379 events from 22 of the tumours were cross-validated by SOLiD long mate pair sequencing. In brief, 5 µg genomic DNA was sheared into ~2 kb fragments (Covaris®S220), circularized with linker sequences, and followed by digestion of the circularized DNA to generate a template used in emulsion PCR. The template was flanked with adaptor sequences, coupled to beads and clonally amplified before immobilizing to a solid surface for 50 bp sequencing run using the SOLiD v4 (Life Technologies). Sequence data was mapped to a reference genome based on the Genome Reference Consortium (http://www.ncbi.nlm.nih.gov/projects/genome/assembly/grc/human/) GRCh37 assembly using Bioscope v1.2.1 software suite (Applied Biosystems). Each sample was sequenced to an average depth of 199x (ranging from 64 to 333) in the tumour, and 195x (ranging from 52 to 308) in the normal.

An additional 64 somatic rearrangements breakpoints from three tumours were verified using amplicon deep sequencing as previously described in Chapter 2 (Quek et al., 2014). In brief, the verification workflow of breakpoints combines the pooling strategy of

amplicons with benchtop sequencing and standard bioinformatics techniques such as batch primer design, de novo assembly of sequencing data, and automated BLAT to confirm the breakpoints of the somatic rearrangements.

3.2.5 Characterization of breakpoint junctions

In order to characterise the breakpoint junctions, the length of homology at breakpoints were calculated based on split contig alignment sequence. The qSV output includes the sequence of any microhomology (overlapping) regions at each breakpoint, it also outputs non-templated sequence, which does not align to the breakpoint junction. The sequences identified as non-templated sequences were analysed using BLAT and then further annotated as genomic shard (DNA fragment which maps elsewhere in the genome) and free DNA with repeat annotation. The following paragraphs described the characteristics of breakpoints junction in more details:

- *Blunt end:* Throughout this chapter, **"**0 bp" represents the blunt end characteristic as there is zero overlapping bases homology at the break. It describes a clean break whereby the broken chromosome ends are re-joined accurately.

- *Microhomology:* This characteristic is presented as "1 to 5 bp" and "6 or + bp" depending on the length of sequence homology at the breakpoint. It describes the presence of terminal short overlapping bases between the two repaired chromosome ends (Chen, 2001). Microhomology length of 1 to 5 bp is the most common feature of NHEJ while length 6 to 25 bp is a feature of MMEJ DNA repair mechanism.

- *Insertion of non-templated sequence:* This characteristic is presented as "-25+ to -1 bp". It describes inserted sequence of free DNA found at the junction of breakpoints. The length of non-templated sequences ranged from approximately 1 to 150 bp based on previous studies on breast and lung cancers (Bignell et al., 2007; Stephens et al., 2009). The non-templated sequence was either poorly or not aligned with reference genomes. The alignment was performed using BLAT to match sequences ≥ 20 bp against human reference genome hg19, USCS database. Such sequence could be generated by enzymatic process during DNA

repair (Stephens et al., 2009) or due to physical shearing of the chromosomes (Bignell et al., 2007).

- *Genomic shard:* The label of "-25+ to -1 bp" includes the genomic shard characteristic. For non-templated sequence that was mapped uniquely to the reference genome near the vicinity of the breakage or elsewhere in the genome are classified as genomic shard. Typically, the length of genomic shard is approximately 20 to 500 bp (Bignell et al., 2007; Campbell et al., 2008). It has been proposed that complex rearrangements involving multiple breakpoints are associated with genomic shards (Hastings et al., 2009). This could be the result of complex rearrangements produced by shattering of chromosomal regions followed by reassembly (chromothripsis) which may result in some large-scale rearrangements containing additional small fragments (genomic shards) inserted at breakpoint (Bignell et al., 2007; Stephens et al., 2011).

- *Free DNA with repeat annotation:* The label of "-25+ to -1 bp" also denotes the characteristic of free DNA with repeat annotation. This characteristic describes non-templated sequence that has overlapped with the sequence of mobile elements.

3.2.6 Analysis of breakpoints characteristics

After the characterization of breakpoints, statistical tests were conducted to evaluate if the patterns of breakpoints characteristics were different between various molecular events. These include genome subtype, BRCA mutational signature and *BRCA* mutation. The mutational signatures are calculated based on the context that somatic mutation occur using the 96 possible combination of bases at 5' and 3' of the mutated base (e.g. C>A, C>G, C>T, T>A, T>C, T>G) across the entire genome (Alexandrov et al., 2013). Statistical tests were performed using the length of homology at the breakpoint junctions of the rearrangements. The statistical tests include (1) unsupervised PCA to search for potential clustering in the analysed dataset and (2) Wilcoxon rank sum test to assess the differences and relationship between various molecular events and cancer types with respect to the categorized length of homology. $P<0.05$ indicates significant result. Tests were performed using Prism and R software (www.r-project.org).

3.2.7 Motif discovery

The MEME motif finder was originally developed to discover sequence motifs such as DNA binding sites and protein interaction domains (Bailey et al., 2009). Here, it was used to find statistically enriched DNA sequence (motifs) surrounding the breakpoint junctions and generate position weight matrices. Two hundred base pairs on either side of the breakpoints were grouped into their respective analysed subsets. Two sequences of 400 bp per rearrangement were used as input for MEME and the default short motif width was set between 5 bp and 8 bp.

3.3 Results

3.3.1 Structural rearrangements in 120 pancreatic primary tumours

Whole genome sequencing enabled us to identify a total of 10,074 high confidence somatic rearrangements from the 120 pancreatic primary tumours with a median of 83 rearrangements per tumour (ranging from 11 to 576) (Figure 3-3). Of 1,379 somatic rearrangements identified from 22 tumours, 59.7% (823) were cross-validated using SOLiD LMP sequencing. An additional 64 rearrangements were also randomly selected from patients APGI 1959, APGI 2049 and APGI 2156 for amplicon deep sequencing. Using the workflow developed in Chapter 2, 73.4% (47 of 64) rearrangements were confirmed as somatic by PCR and 70.2% (33 of 47) had their exact breakpoint location identified by aligning MiSeq sequencing results to the human reference genome (Supplementary Table 3-1).

The frequency of different types of somatic rearrangements varied among individuals across the cohort. The somatic rearrangements were classified as intra-chromosomal (events within a single chromosome) or inter-chromosomal (events between two different chromosomes, i.e. translocations). The proportion of intra-chromosomal rearrangements in this dataset was 87.8% (8,847 of 10,074) and inter-chromosomal rearrangements was 12.2% (1,227 of 10,074) highlighting that intra-chromosomal rearrangements generally prevail in pancreatic tumours. This is consistent to what have been described for 13 previously characterised pancreatic tumours (Campbell et al., 2010) and is also similar to other cancer types including colorectal, breast and liver (Kloosterman et al., 2011; Stephens et al., 2009; Totoki et al., 2011). The orientation of the read pairs and the association of copy number change enables intra-chromosomal events to be further classified as inversions, amplified inversions, foldback inversions, deletions, duplications or tandem duplications. Events that are not associated with an inversion or a copy number change were broadly grouped as other intra-chromosomal rearrangements. The distribution of somatic rearrangements was as follows: 14.7% (1,481) inversions, 4.2% (426) amplified inversions, 6.4% (647) foldback inversions, 12.8% (1,291) deletions, 12.2% (1,227) translocations, 1.3% (130) duplications, 1.7% (175) tandem duplications, 46.6% (4,697) other intra-chromosomal rearrangements (Figure 3-3). Inversions formed the most common sub-classification, which is in contrast to primary breast cancer genomes, whereby tandem duplications are the most common sub-classification (Stephens et al., 2009).

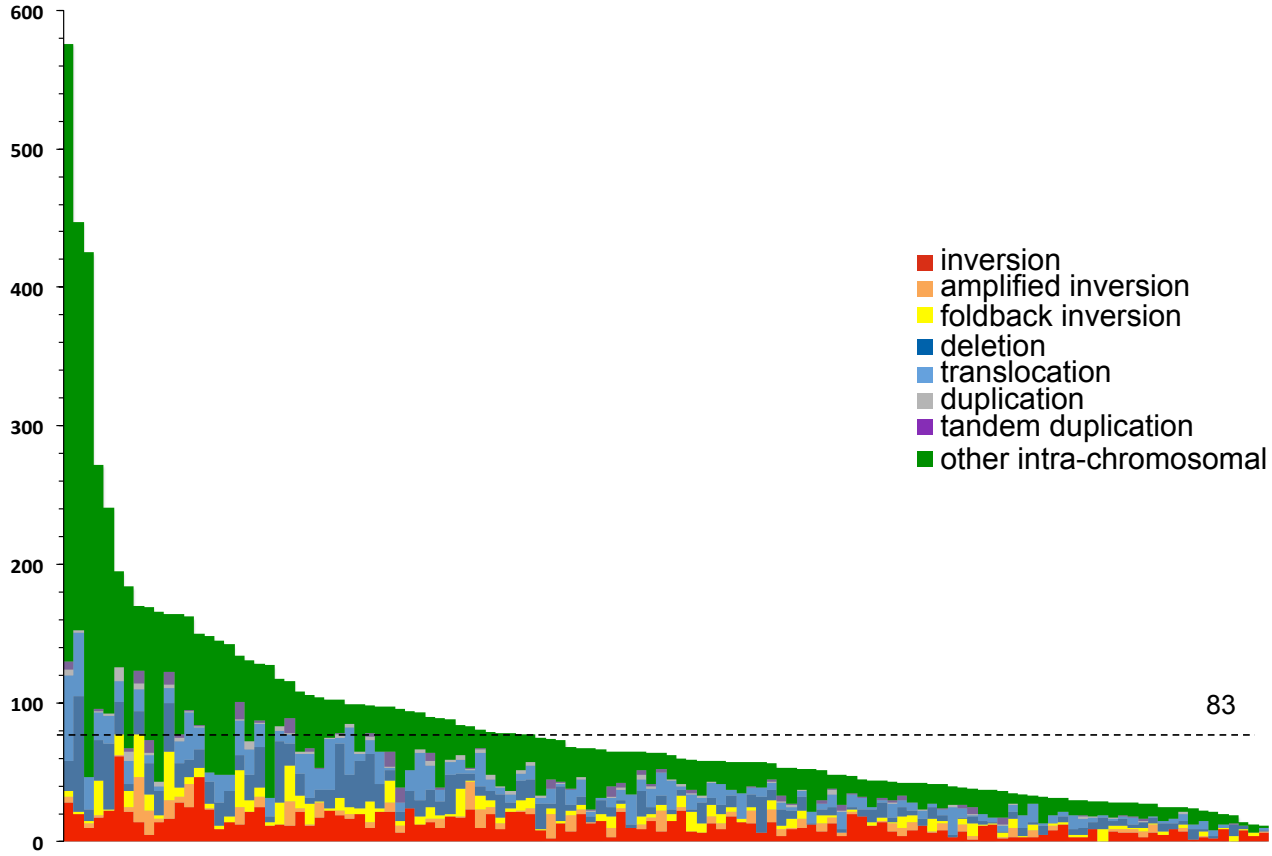


Figure 3-3 Spectrum of somatic rearrangements and different types across 120 pancreatic primary tumours. Patients are presented on the x-axis and the number of each rearrangement types is on the y-axis. Rearrangements are coloured according to the event types as shown in the legend. Dashed line indicates the median of rearrangements per tumour.

3.3.2 Characterisation of breakpoints

The formation of somatic rearrangements in cancer genomes may rely on sequence homology dependent and/or independent mechanisms (Chen, 2001; Korbel et al., 2007; Lam et al., 2010). Here, I characterised the breakpoints junctions as blunt end, microhomology, non-templated insertion, genomic shard or free DNA with repeat annotation. Of the 10,074 high confidence somatic rearrangements, a total of 9,741 contained split contig alignment, which allowed the prediction of the sequence context at the rearrangements junction to facilitate the characterization of breakpoint phenotypes (Figure 3-3). The proportion of these breakpoints characteristics within the pancreatic cohort are shown in Figure 3-4 and were characterised as:

Blunt end: Of 9,741 somatic rearrangements, 1,207 (12.4%) were blunt end with an average of 10 rearrangements per patient (ranging from 0 to 56). Two of the 120 patients did not display blunt end characteristics (APGI 2598 and APGI 3513).

Microhomology: In this dataset, the vast majority of the rearrangements shared short stretches of overlapping bases. A total of 7,283 (74.8%) rearrangements displayed microhomology with an average of 61 rearrangements per patient (ranging from 7 to 448). No long homologous sequences (>100 bp) were observed in this dataset (the longest homology sequence seen was 17 bp long).

Insertion of non-templated sequences: Of 9,741 rearrangements, 1,251 (12.8%) displayed insertion of free DNA at the breakpoint junctions. Of these, 1,160 (12.0%) displayed non-templated sequence insertion characteristic at the breakpoints junctions with an average length of 6.4 bp (ranging from 1 to 68 bp) (Figure 3-4). These non-templated sequences were either poorly or not aligned with reference genomes.

Genomic shard: Thirty-nine (0.4%) of the rearrangements were classified as genomic shard with an average length of 34.2 bp (ranging from 20 to 76 bp). Genomic shards were quite uncommon and observed in 29 of 120 pancreatic tumours.

Free DNA with repeat annotation: Fifty-two (0.5%) rearrangements had free DNA situated along mobile elements such as LINEs, SINEs, LTR retrotransposons and DNA repeats. The average length of these free DNA was 34.0 bp (ranging from 20 to 71 bp) and this characteristic was observed in 37 of 120 pancreatic tumours.

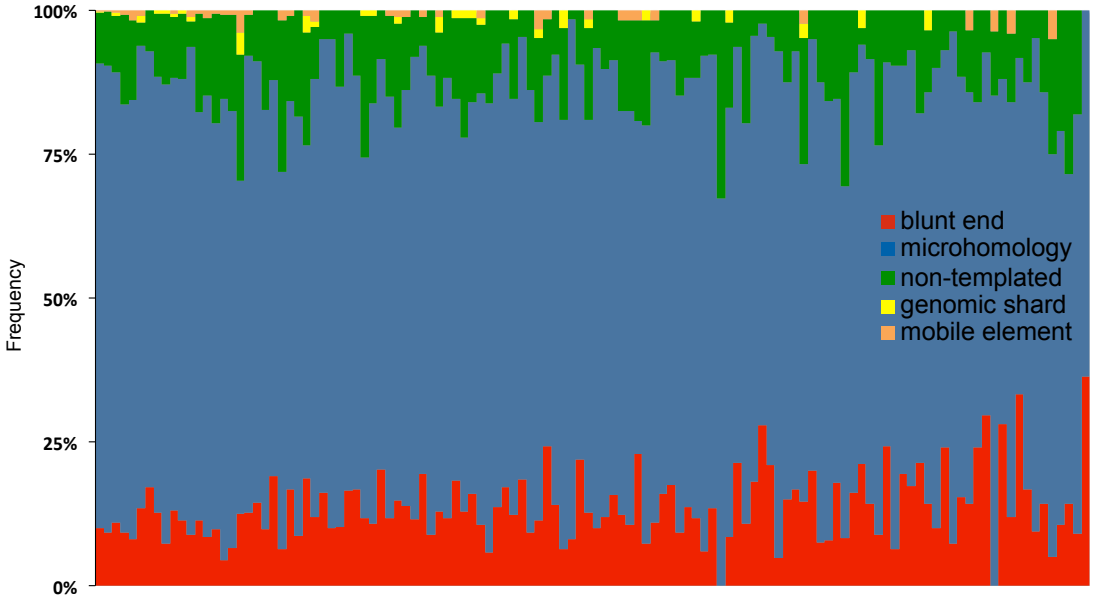


Figure 3-4 Proportion of breakpoints characteristics across 120 pancreatic primary tumours. Patients are on the x-axis and the percentage of the breakpoints characteristics for each rearrangement within each tumour is on y-axis.

In summary, these characteristics of breakpoints (i.e. sequence homology or sequence insertion) may represent "scars" in the cancer genome displaying their history. The estimated contributions of breakpoint characteristics are consistent to what have been reported in previous studies with microhomology being the largest contributor of all characteristics (Ng et al., 2012; Stephens et al., 2011) (Table 3-1).

Table 3-1 Characteristics of somatic rearrangements breakpoints

Type of characteristics	Estimated contribution (n=120)	Contribution reported by Ng et al, 2012 (n= 2)	Contribution reported by Stephens et al, 2011 (n = 11)	Possible mechanism(s) associated with somatic rearrangements formation
Blunt ends (0 bp)	12.4%	16.0%	23.1%	'accurate' NHEJ
Microhomology	74.8%	68.0%	56.5%	'error-prone' NHEJ, MMEJ
Non-templated sequence (<20 bp)	12.0%	16.0%	13.2%	'error-prone' NHEJ
Genomic shard (insertion of non-templated sequence (>20 bp)	~0.4%	0%	7.2%	'error-prone' NHEJ, chromothripsis
Mobile element (≥20 bp)	~0.5%	-	-	Retrotransposition

3.3.3 Microhomology of somatic rearrangements

According to the literature, the different lengths of short overlapping bases at breakpoints reveals the involvement of non-homologous recombination mechanism, which is also known to have two sub-pathways: (1) non-homologous end joining (NHEJ) which uses 1-5 bp microhomology and (2) microhomology-mediated end joining (MMEJ) uses >5 bp microhomology (Chen, 2001; Hastings et al., 2009).

Of 9,741, 6,525 (67.0%) rearrangements contained microhomology of 1-5 bp suggesting that NHEJ is more frequent in pancreatic cancer genomes while the prevalence of MMEJ is approximately 7.8% (758 rearrangements had >5 bp microhomology) (Figure 3-5). Overall, the rearrangement breakpoints showed an average length of 3.0 bp microhomology (ranging from 1 to 17 bp). When comparing to the somatic rearrangements detected in 95 tumours (such as breast, head and neck, colorectal, prostate carcinomas, melanoma, multiple myeloma, and chronic lymphocytic leukaemia) (Drier et al., 2013), our cohort of pancreatic cancer genomes has longer microhomology length at breakpoints junctions (average of 3.0 bp vs. 1.7 bp). This suggests that the efficiency of DNA repair in our pancreatic cohort might be higher than the cohort studied by Drier et al. as the efficiency of DNA repair has been shown to increase as the length of homology increases (Villarreal et al., 2012).

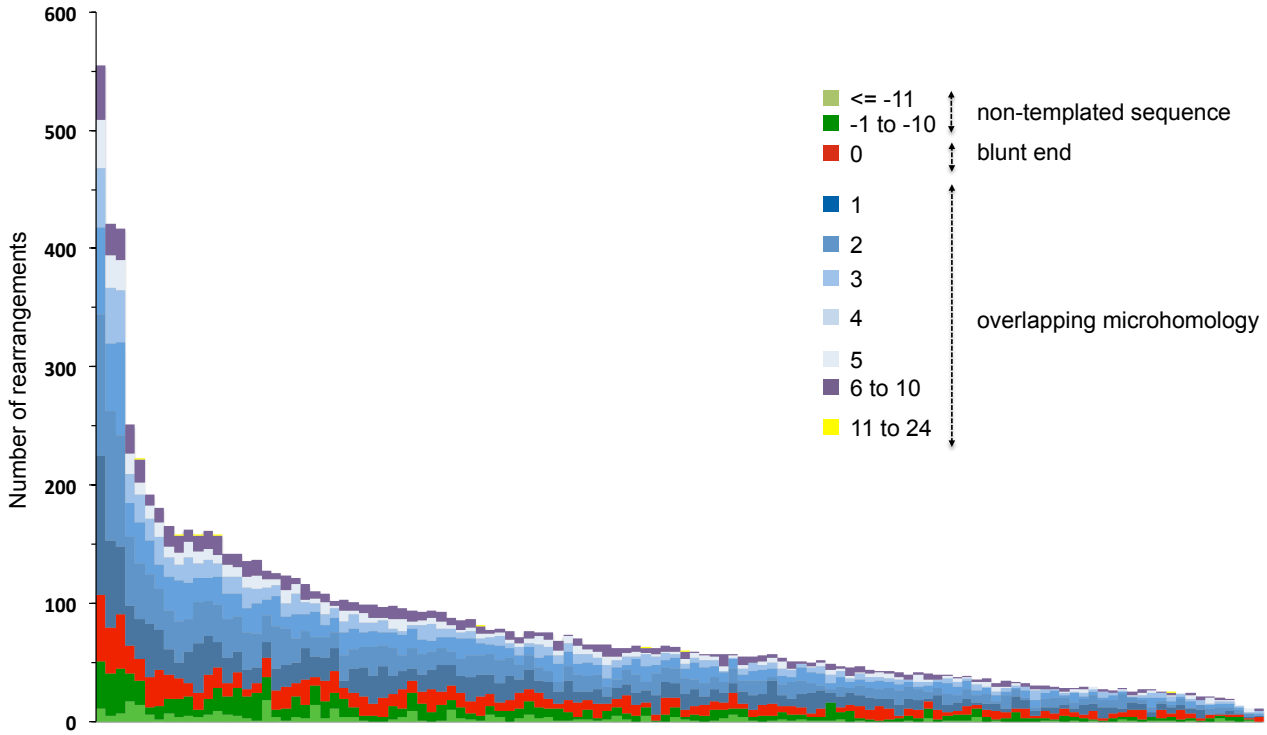


Figure 3-5 Breakpoint characteristics across 120 pancreatic primary tumours. Microhomology in bp are shown for each breakpoints as a positive number. Patients on the x-axis are ordered by the number of somatic rearrangements (y-axis). Blunt end has homology of 0 bp. Non-templated sequences (insertion of free DNA) are shown as negative numbers.

Drier et al. reported that the distribution of microhomology lengths varied by the type of rearrangement across the 95 tumours (Drier et al., 2013). To determine if different event types (deletions, inversions, duplications, translocations and intra-chromosomal rearrangements) show different breakpoint characteristics and thus potentially are repaired differently in pancreatic cancer, the distribution of microhomology lengths for each type was analysed. (Figure 3-6). In our dataset, the distributions of microhomology length are indeed varied by different event types which agree with what was reported in Drier et al. (Drier et al., 2013). It was observed that microhomology of 1 or 2 bp is the most common overlapped sequence across majority of the event types (Figure 3-6). This agrees with the observations in breast cancers, where most of the rearrangement types had distinctive distribution with 2 bp except in amplification, which blunt end (0 bp) was the most common (Stephens et al., 2009).

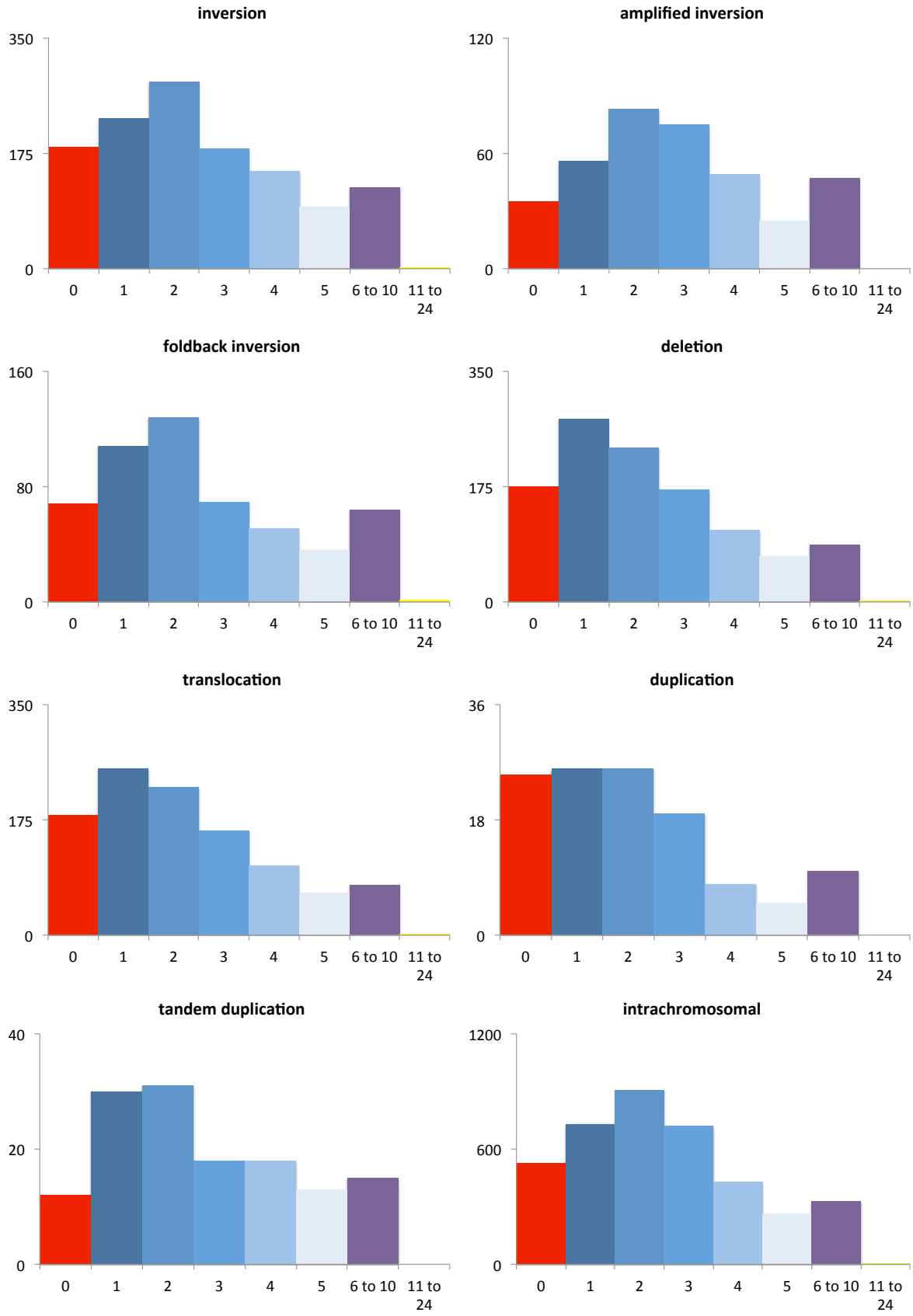


Figure 3-6 Distribution of microhomology length of somatic breakpoints by event type across 120 pancreatic tumours. The number of somatic rearrangements is plotted against the number of base pairs of microhomologies. The length of microhomologies is presented on the x-axis and the number of rearrangements is on the y-axis.

3.3.4 Characteristics of breakpoints in cancer genomes

Members of ICGC Australia pancreatic team recently grouped pancreatic tumours (PDAC) into 4 subtypes termed stable, scattered, focal and unstable (Figure 3-7) (Biankin et al., 2012). The unstable tumours were found to be associated with genome instability and contained a large number of rearrangements. Many, but not all of the unstable tumours contained mutations in key DNA damage genes (*BRCA*1/2, *ATM* or *PALB*2) and were associated with a high BRCA mutational signature. Patients with an unstable genome also responded to platinum-based chemotherapies probably due to the defective HR pathway. The identification of which tumours harbour a defective HR pathway is clinically important, however, the unstable subtype or *BRCA* mutation alone is not able to predict a defective HR pathway with 100% accuracy. Here, I tested whether breakpoint characteristics in 120 pancreatic tumours may indicate which tumours have a defective *BRCA* pathway. Initially, the analysis was performed to look for difference in breakpoint characteristics between the genomic subtypes, then between *BRCA* carriers and finally between tumours with high or low BRCA signatures. In an attempt to further explore and validate findings, data from ovarian (AOCS) and oesophageal cancers (OESO) were incorporated in a cross cancers analysis.

3.3.4.1 Breakpoints characteristics in the different genome subtypes of PDAC

Here, I evaluated if there are differences in breakpoints characteristics across the 4 genome subtypes that could identify potential mechanisms involved in the formation of somatic rearrangements. The cohort of 120 PDAC were grouped as: 32 focal (26.7%), 64 scattered (53.3%), 15 stable (12.5%), and 9 unstable (7.5%) genome subtypes.

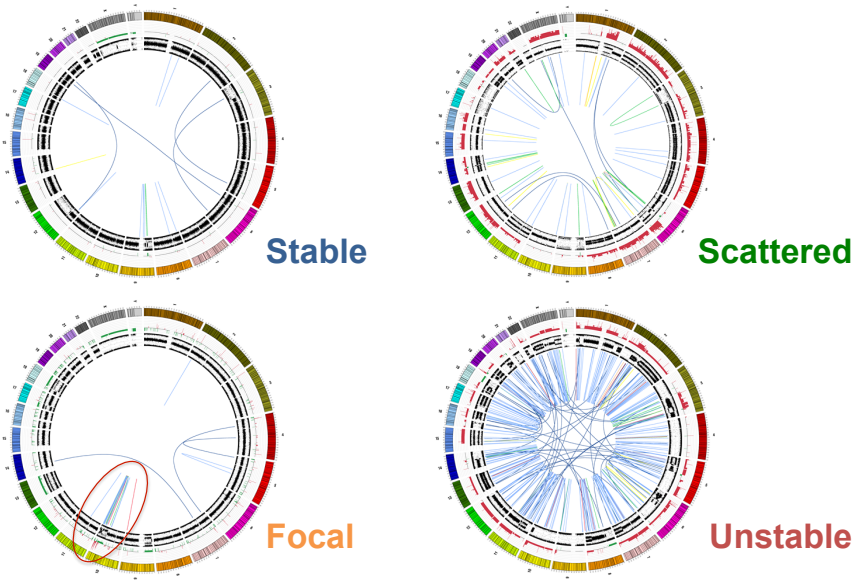


Figure 3-7 Classification of PDAC primary tumours based on the genomic profile. A representative tumour from each subtype is shown using Circos plots (Krzywinski et al., 2009). Chromosome ideograms are shown in the outer ring. The inner two rings show the copy number profiles obtained by GAP analysis (Genome Alteration Print) (Popova et al., 2009) and B allele frequency. Within the circle, each line connects the sequenced junctions of rearrangements – translocation (dark blue), deletions (green), inversions (yellow), tandem duplications (dark red), duplications (red), foldback inversions (light orange), amplified inversions (dark orange), and intra-chromosomal (very light blue). The numbers and the distribution of somatic rearrangements within a genome were used to classify tumours into 4 subtypes (Waddell et al., 2015)

The number of events with different breakpoint characteristics was used in the PCA analysis to identify subgroups. PCA suggests that there was difference between the unstable and stable group (Figure 3-8). However, this difference might be associated with the number of events since the genome subtypes were originally based on the number of events in the genome. Breakpoint characteristics were not able to separate tumours within focal subtype. This is perhaps due to the peculiar phenotype of the focal subtype where most of the breakpoints were localized on a single chromosome and contained features of complex rearrangements including chromothripsis and breakage-fusion-bridge. In contrast, tumours with the scattered and stable subtypes contained events which are randomly distributed through the genome.

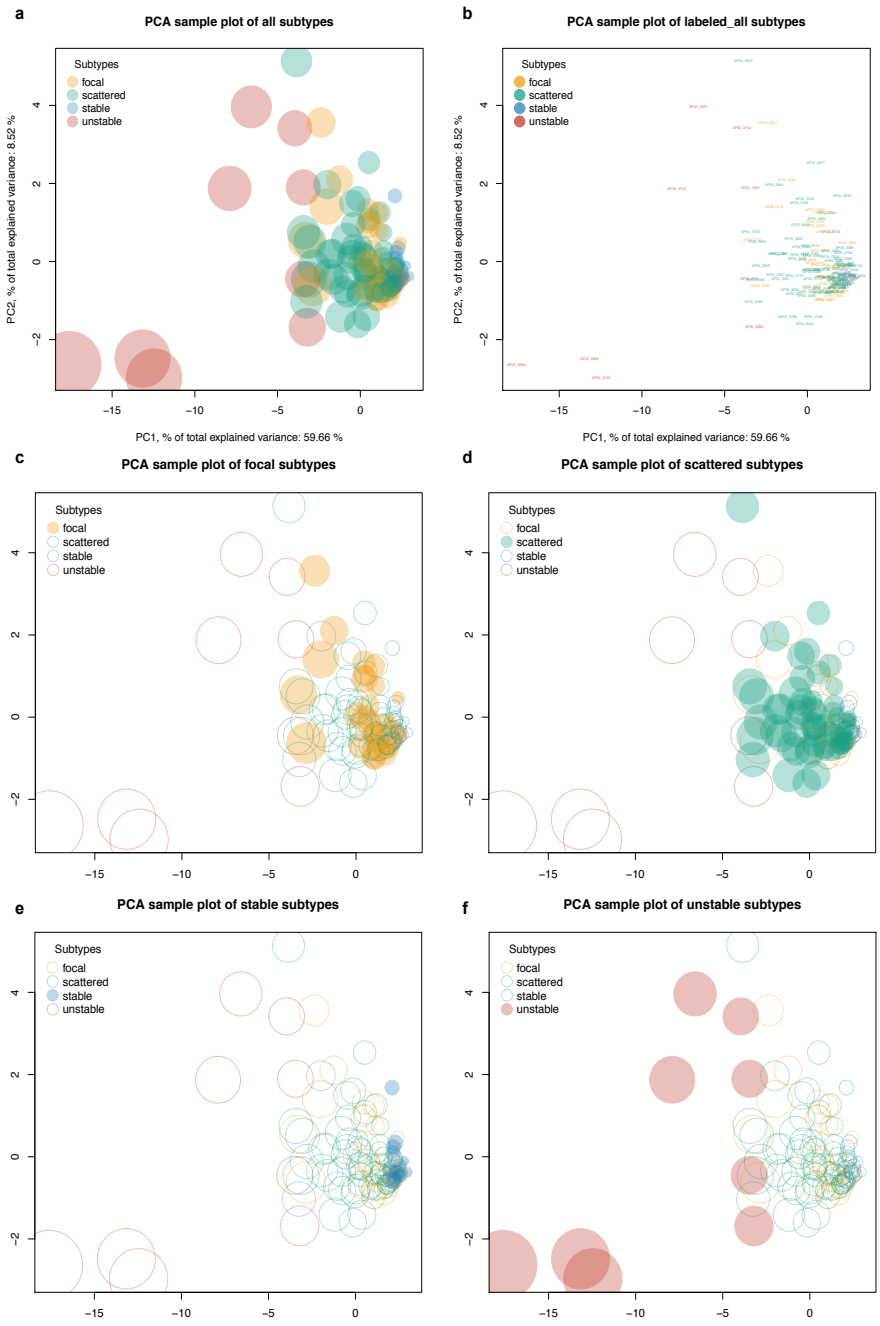


Figure 3-8 PCA plot of breakpoints across the 4 genome subtypes of PDAC. Each bubble represents a pancreatic cancer sample. Samples are coloured according to the genome subtypes – focal, orange; scattered, green; stable, blue; unstable, red. Size of the bubble corresponds to the number of somatic rearrangements in each sample. (a) PCA was applied to all 9,741 somatic rearrangements with measured length of homology for 120 samples to calculate the variance of the principal components. PCA was plotted in two dimensions using their projections onto the first two principle components. (b) The samples were labelled with unique patient ID. (c) PCA plot emphasized on focal subtype. (d) PCA plot emphasized on scattered subtype. (e) PCA plot emphasized on stable subtype. (f) PCA plot emphasized on unstable subtype.

To determine whether the genome subtypes are associated with specific repair pathways, the frequency distribution of different length of homology across the 4 genome subtypes (i.e. focal, scattered, stable and unstable) was assessed. The unstable subtype which was associated with the HR defective pathway, showed a lower frequency of breakpoint with blunt end characteristic suggesting that the HR pathway might not be intact. A significance difference between scattered and unstable subtypes was observed in blunt end characteristic (p=0.0422, Wilcoxon rank-sum test) (Figure 3-9) (Chen, 2001). However, the distribution of categorized length of homology between stable and unstable subtypes showed no differences suggesting that the difference observed between stable and unstable subtypes observed in PCA analysis could be driven by the number of rearrangements in a sample as the unstable tumours have a higher number of somatic rearrangements.

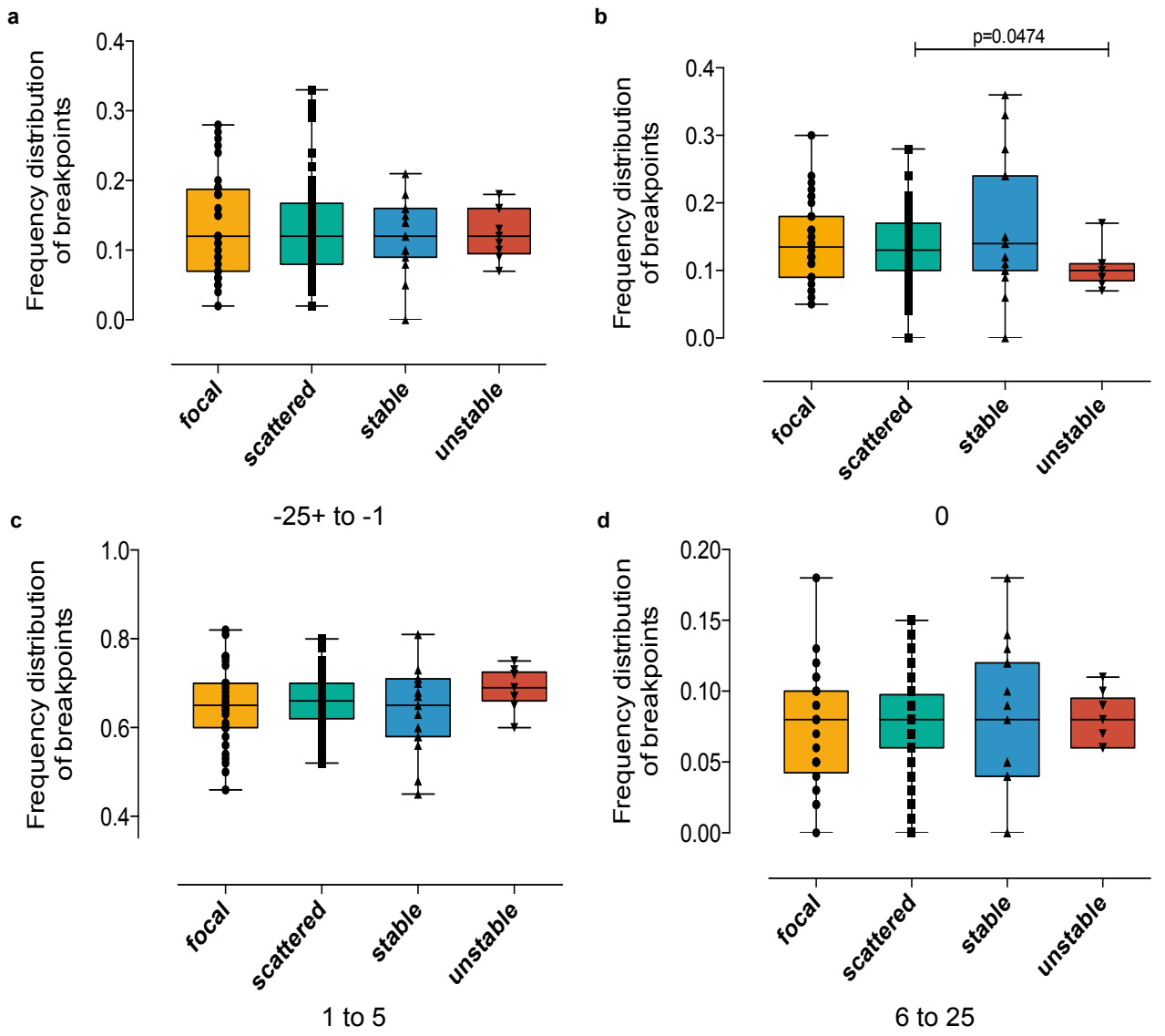


Figure 3-9 Frequency distribution of breakpoints across 4 genome subtypes of PDAC. Each genome subtype (x-axis) was plotted against the frequency distribution of somatic rearrangements (y-axis). The sample size for each genome subtype was as follows: focal (n=32), scattered (n=64), stable (n=15), and unstable (n=9). Boxplots were coloured according to the genome subtype – scattered, green; stable, blue; unstable, red. (a) The categorized length of homology -25+ to -1 bp denotes the characteristic of free DNA insertion at the break. (b) The categorized length of homology 0 bp denotes the characteristic of blunt end. (c) and (d) The categorized length of homology 1 to 5 bp and 6 to 25 bp denote the characteristic of microhomology. All p-values are from Wilcoxon rank-sum test. Comparisons with p-value >0.05 were not labelled.

3.3.4.2 Breakpoint characteristics associated with BRCA mutational signature of PDAC

The sequence context of somatic substitutions and indels has been shown to predict the likely mutation processes or signatures within the different tumours. Twenty-one mutational signatures were initially described from the analysis of 7,042 tumours from 30 different cancer types (Alexandrov et al., 2013). These mutational signatures were associated with biological features or characteristic related to cancer, for example, age, smoking APOBEC, *BRCA*1/2 mutations, DNA MMR deficiency, ultraviolet light, immunoglobulin gene hypermutation, Pol ε mutations, and temozolomide. The signatures which the authors termed, the "BRCA signature" was associated with inactivating mutations of the *BRCA*1/2 genes. Tumours with *BRCA*1 and *BRCA*2 mutations showed a large contribution from *BRCA* mutational signature (reported as signature 3). However, some tumours with a substantial contribution from *BRCA* signature did not have *BRCA*1 and *BRCA2* mutations, indicating that other mechanisms or genes rather than solely *BRCA*1 and *BRCA2* inactivation generate this signature (Alexandrov et al., 2013).

In pancreatic cancer, the presence of mutation in the *BRCA*1/2 genes and high BRCA mutational signature were associated with the unstable subtype of pancreatic cancer. However, not all tumours with a high BRCA mutational signature harboured *BRCA* gene mutations and not all unstable tumours contained a high BRCA signature.

To determine if the BRCA signature was associated with defects in specific DNA repair pathways, the differences in breakpoints characteristics of the tumours with a high and low contribution of *BRCA* mutational signature were evaluated. The number of

mutation per Mb associated with BRCA mutational signature for 120 pancreatic primary tumours ranged from 0.00 to 0.70 (Figure 3-10). To evaluate the breakpoint characteristics, the upper quartile (i.e. high contribution of BRCA signature) of the BRCA mutational signature was compared to the lower quartile (i.e. low contribution of BRCA signature). The value of high BRCA mutational signature ranged from 0.41 to 0.70 and the value of low BRCA mutational signature ranged from 0 to 0.14. A total of 14 tumours with a high contribution of the BRCA mutational signature and 30 tumours with a low contribution of the BRCA signature were compared.

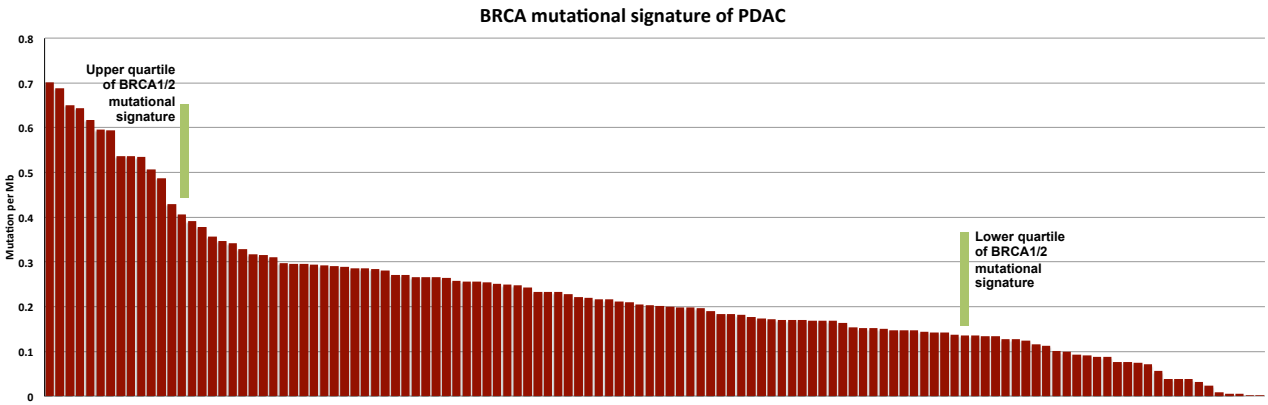


Figure 3-10 The number of mutation per Mb that contributed for BRCA mutational signature within each PDAC sample. The samples were ranked by prevalence (red bars). The green bar marks the cut-off for high and low contribution of *BRCA* signature.

There was a difference in the characteristics of breakpoints for blunt end (length 0 bp; p=0.0137, Wilcoxon rank-sum test) and microhomology (length 1 to 5 bp; p=0.0487, Wilcoxon rank-sum test) between tumours with a high and low BRCA signature (Figure 3-11). These results suggest that the two known NHEJ pathways 'accurate' and 'error-prone' (Pfeiffer et al., 2000) might have different frequency of activity in tumours with high and low BRCA mutational signature. The 'accurate' and 'error-prone' NHEJ share different elements of HR pathway (Pfeiffer et al., 2000), thus the 'accurate' NHEJ is independent of homology sequence resulting in blunt end ligation (0 bp) while the 'error-prone' NHEJ uses microhomology to repair non-complementary ends. The results here suggest that the tumours with a high BRCA mutational signature might present a HR deficiency pathway with less activity of the 'accurate' NHEJ. These BRCA deficiency tumours seem to favour the 'error-prone' NHEJ pathway that uses microhomology to repair the chromosomal breaks.

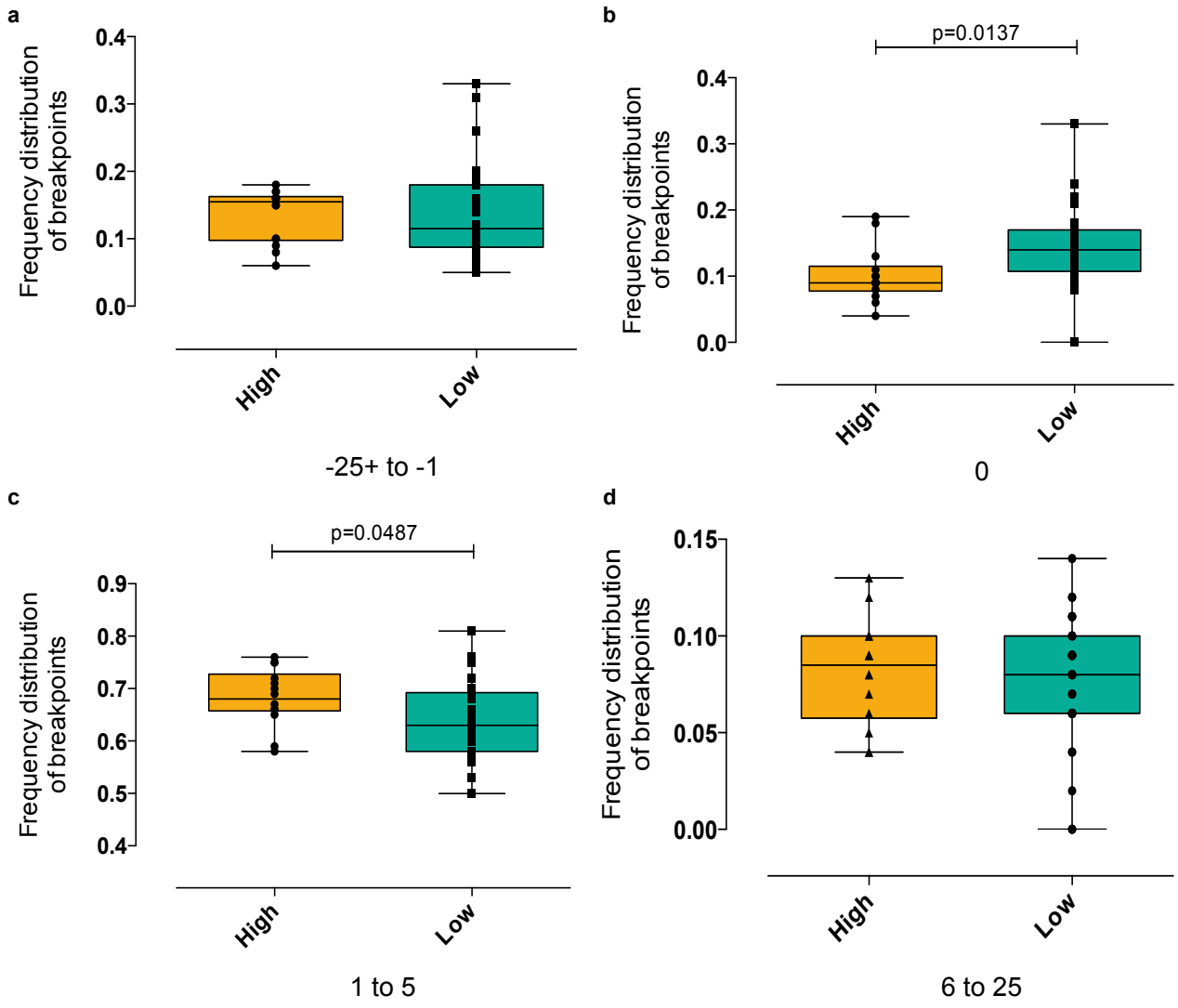


Figure 3-11 Frequency distribution of breakpoints of BRCA mutational signature in PDAC samples. The number of mutations per Mb that contributed to the BRCA mutational signature was calculated. The samples within the high and low BRCA mutational signature cut-off value (x-axis) were plotted against the frequency of somatic rearrangements (y-axis). The sample size for a high and low BRCA mutational signature was as follows: high (n=14) and low (n=30). Boxplots were coloured according to quartiles of BRCA mutational signature – high, orange; low, green. (a) The categorized length of homology -25+ to -1 bp denotes the characteristic of free DNA insertion at the break. (b) The categorized length of homology 0 bp denotes the characteristic of blunt end. (c) and (d) The categorized length of homology 1 to 5 bp and 6 to 25 bp denote the characteristic of microhomology. All p-values are from Wilcoxon rank-sum test. Comparisons with p-value >0.05 were not labelled.

3.3.4.3 Breakpoint characteristics associated with *BRCA* mutation of PDAC

As tumours with a high *BRCA* mutational signature do not always have *BRCA* mutations, the breakpoint characteristics of tumours that were BRCA mutant and *BRCA* wild type were investigated. A total of 6 tumours with *BRCA* mutation were compared with 114 tumours with *BRCA* wild type.

The frequency distribution of blunt end breakpoints (0 bp) in tumours with *BRCA* mutation was significantly lower as compared to *BRCA* wild type tumours (p=0.0189, Wilcoxon rank-sum test) (Figure 3-12). In agreement with the previous results observed in the BRCA mutational signature, here the frequency of breakpoints with categorized length of homology 1 to 5 bp was higher in tumours with *BRCA* mutation than in tumours with *BRCA* wild type. However, the difference was not statistically significant. As there were only 6 tumours with *BRCA* mutation, this analysis might be limited by the sample size.

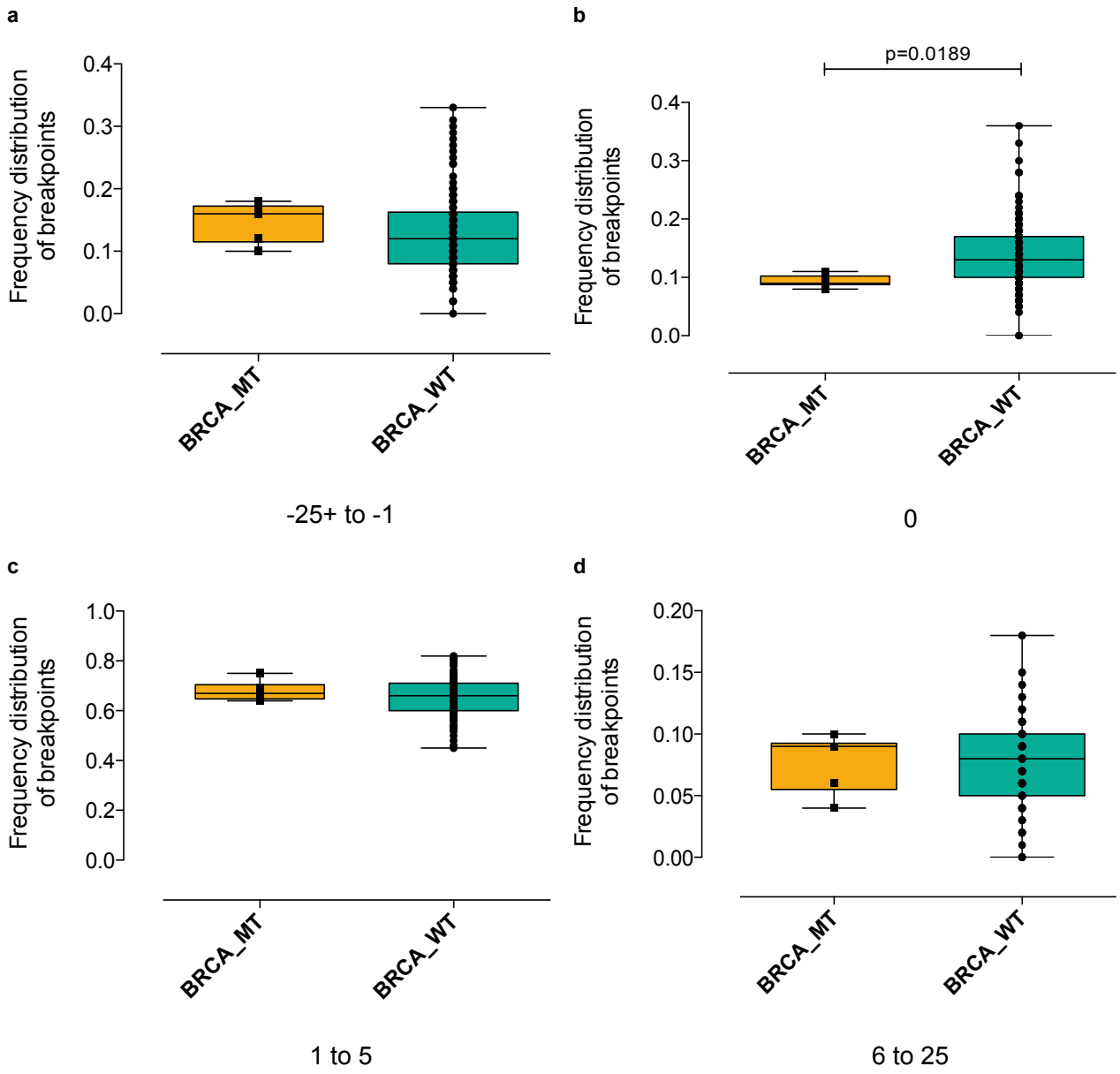


Figure 3-12 Frequency distribution of breakpoints of *BRCA* gene mutation in PDAC samples. *BRCA* mutation status was plotted against the frequency of somatic rearrangements (y-axis). The sample size for each mutation status was as follows: BRCA_MT (n=6), and BRCA_WT (n=114). Boxplots were coloured according to *BRCA* gene mutation status – samples with *BRCA*1/2 mutant (BRCA_MT), orange; samples with *BRCA*1/2 wild type (BRCA_WT), green. (a) The categorized length of homology -25+ to -1 bp denotes the characteristic of free DNA insertion at the break. (b) The categorized length of homology 0 bp denotes the characteristic of blunt end. (c) and (d) The categorized length of homology 1 to 5 bp and 6 to 25 bp denote the characteristic of microhomology. All p-values are from Wilcoxon rank-sum test. Comparisons with p-value >0.05 were not labelled.

3.3.4.4 Breakpoint characteristics associated with BRCA mutational signature of AOCS

To verify the association with BRCA mutational signature and breakpoint characteristics, data from ovarian cancer was analysed, as the frequency of *BRCA* gene mutations are higher in ovarian cancer genomes. Therefore, in ovarian cancer, I further investigated the difference in characteristics of breakpoints for the following: (1) BRCA mutational signature and (2) *BRCA* gene mutation in ovarian cancer, the analysis was conducted in 80 primary high-grade serous ovarian cancer genomes.

The contribution of the BRCA mutational signature in mutation per Mb for ovarian cancer genomes ranged from 0.034 to 4.32 (Figure 3-13), which is much higher than pancreatic cancer (ranged from 0.00 to 0.70). To evaluate the breakpoint characteristics, the upper quartile (i.e. high contribution of BRCA signature) of the BRCA mutational signature was then compared to the lower quartile (i.e. low contribution of BRCA signature). The value of high BRCA mutational signature ranged from 1.82 to 4.32 and low BRCA signature ranged from 0.03 to 0.79. A total of 14 tumours with a high BRCA mutational signature were compared with 29 tumours containing a low contribution of the BRCA mutational signature.

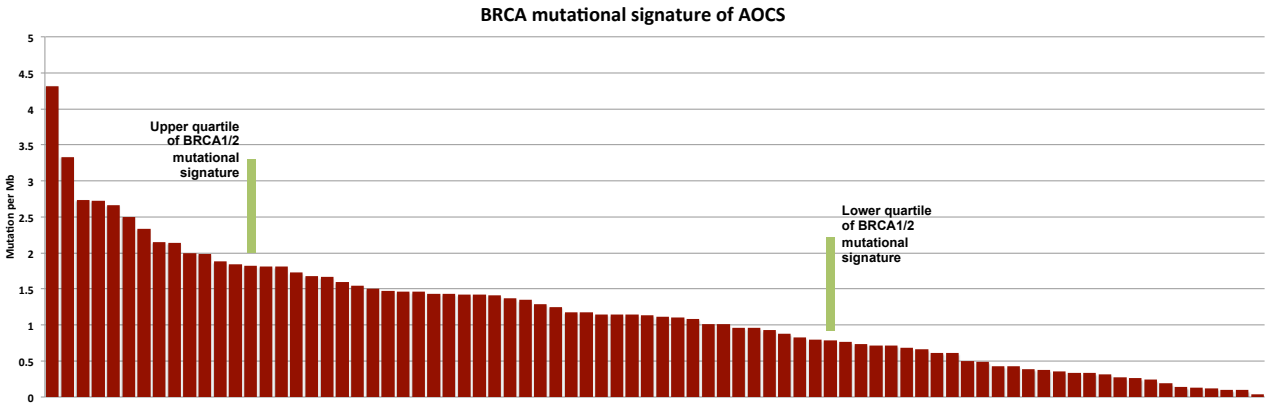


Figure 3-13 The number of mutation per Mb that contributed for BRCA mutational signature within each ovarian sample. The samples were ranked by prevalence (red bars). The green bar marks the upper quartile (i.e. high) and lower quartile (i.e. low) of BRCA mutational signature.

The frequency distribution of breakpoint characteristics showed a similar trend when comparing to the results obtained from pancreatic cancer. Tumours with high BRCA

signature had fewer somatic breakpoints displaying blunt end (0 bp) and more somatic breakpoints displaying microhomology (1 to 5bp) (p=0.0463, Wilcoxon rank-sum test) suggesting again that 'error-prone' NHEJ could be more active in HR deficient tumours as suggested by Pfeiffer et al. (Pfeiffer et al., 2000) (Figure 3-14).

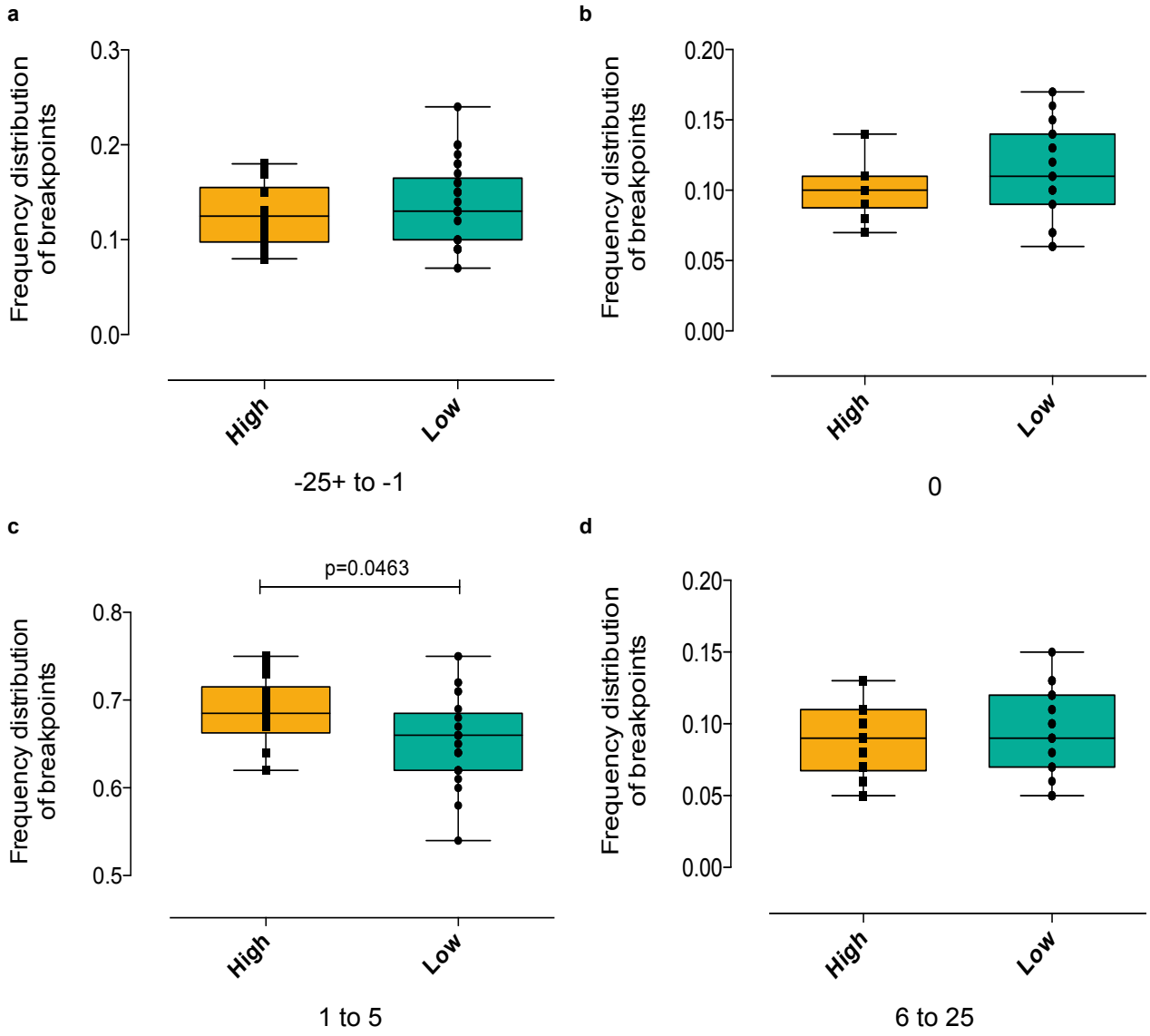


Figure 3-14 Frequency distribution of breakpoints of BRCA mutational signature in AOCS samples. The number of mutations per Mb contributed to the BRCA mutational signature was calculated. The samples within the high and low BRCA mutational signature cut-off value were plotted against the frequency of somatic rearrangements (y-axis). The sample size for a high and low BRCA mutational signature was as follows: high (n=14) and low (n=29). Boxplots were coloured according to quartiles of BRCA mutational signature – high, orange; low, green. (a) The categorized length of homology -25+ to -1 bp denotes the characteristic of free DNA insertion at the break. (b) The categorized length of homology 0 bp denotes the characteristic of blunt end. (c) and (d) The categorized length

of homology 1 to 5 bp and 6 to 25 bp denote the characteristic of microhomology. All p-values are from Wilcoxon rank-sum test. Comparisons with p-value >0.05 were not labelled.

3.3.4.5 Breakpoint characteristics associated with *BRCA* mutation of AOCS

A total of 33 ovarian tumours with *BRCA* mutation were compared to 25 tumours with *BRCA* wild type. However, similar to the *BRCA* signature, the frequency distribution showed that blunt end (0 bp) and microhomology (1 to 5 bp) breakpoints were differed between tumours with *BRCA* mutation and *BRCA* wild type. Expectedly, the tumours with *BRCA* mutation had higher frequency of microhomology breakpoints compared to *BRCA* wild type tumours while tumours with *BRCA* wild type had higher frequency of blunt end breakpoints (blunt end: p=0.0147, microhomology: p=0.0031, Wilcoxon rank-sum test) (Figure 3-15).

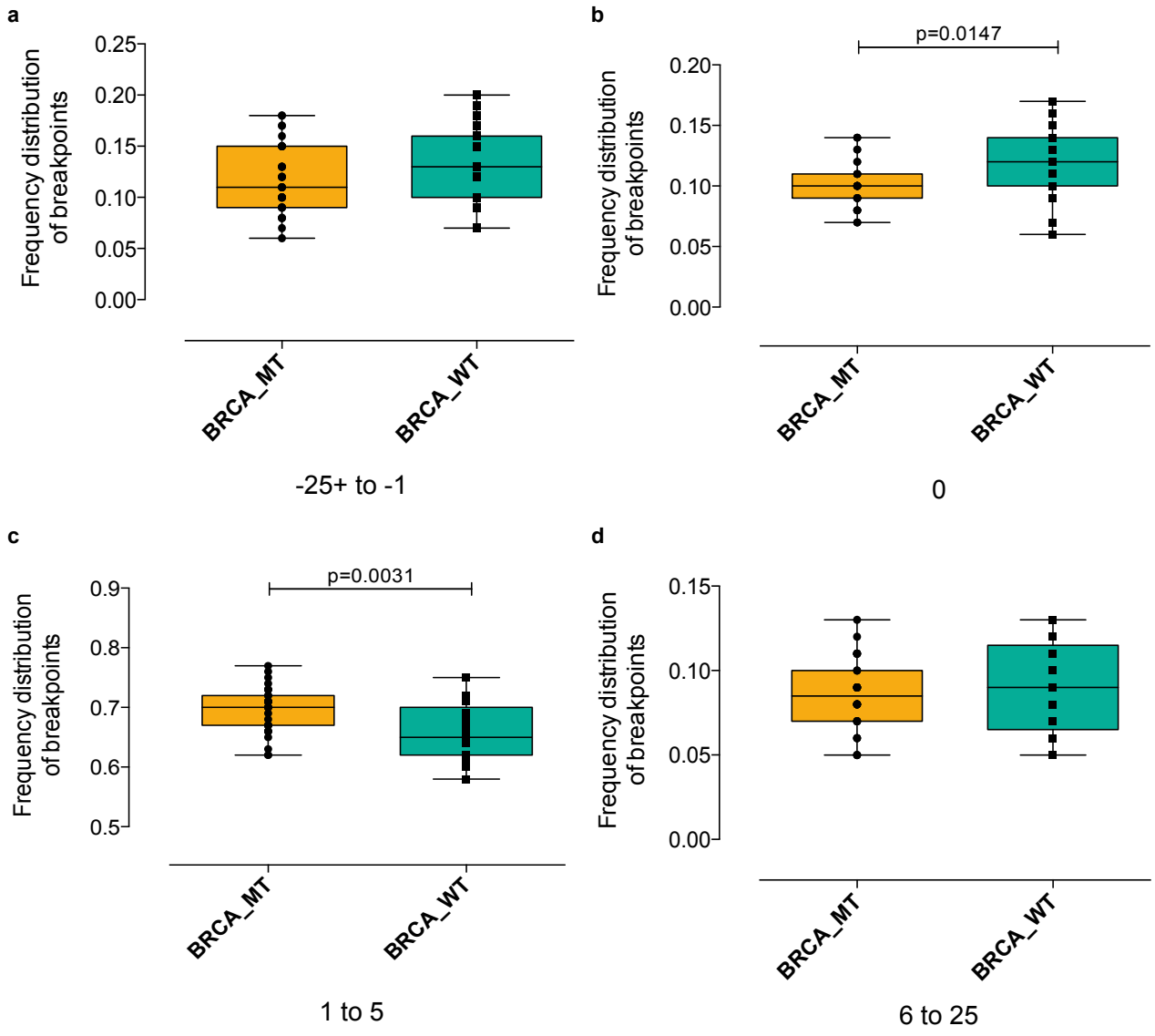


Figure 3-15 Frequency distribution of breakpoints of BRCA gene mutation in AOCS samples. *BRCA* mutation status was plotted against the frequency of somatic rearrangements (y-axis). The sample size for each mutation status was as follows: BRCA_MT (n=33) and BRCA_WT (n=25). Boxplots were coloured according to *BRCA* gene mutation status – samples with *BRCA* mutant (BRCA_MT), orange; samples with *BRCA* wild type (BRCA_WT), green. (a) The categorized length of homology -25+ to -1 bp denotes the characteristic of free DNA insertion at the break. (b) The categorized length of homology 0 bp denotes the characteristic of blunt end. (c) and (d) The categorized length of homology 1 to 5 bp and 6 to 25 bp denote the characteristic of microhomology. All p-values are from Wilcoxon rank-sum test. Boxplots with p-value >0.05 were not labelled.

3.3.4.6 Breakpoint characteristics associated with *BRCA* gene mutation across different tumour types

To increase the number of tumours with *BRCA* mutation, samples in PDAC, AOCS and OESO datasets were combined. Figure 3-16 shows that the distribution of breakpoints characteristics in 40 *BRCA* mutant (6 PDAC and 33 AOCS and 1 OESO) and 182 *BRCA* wild type tumours (114 PDAC, 47 AOCS and 21 OESO). Consistent with previous sections, the frequency of breakpoints with blunt end (0 bp) was higher in tumours with *BRCA* wild type while the frequency of breakpoints with microhomology (1 to 5 bp) was higher in tumours with *BRCA* mutation (blunt end: p=0.0002, microhomology: p<0.0001, Wilcoxon rank-sum test).

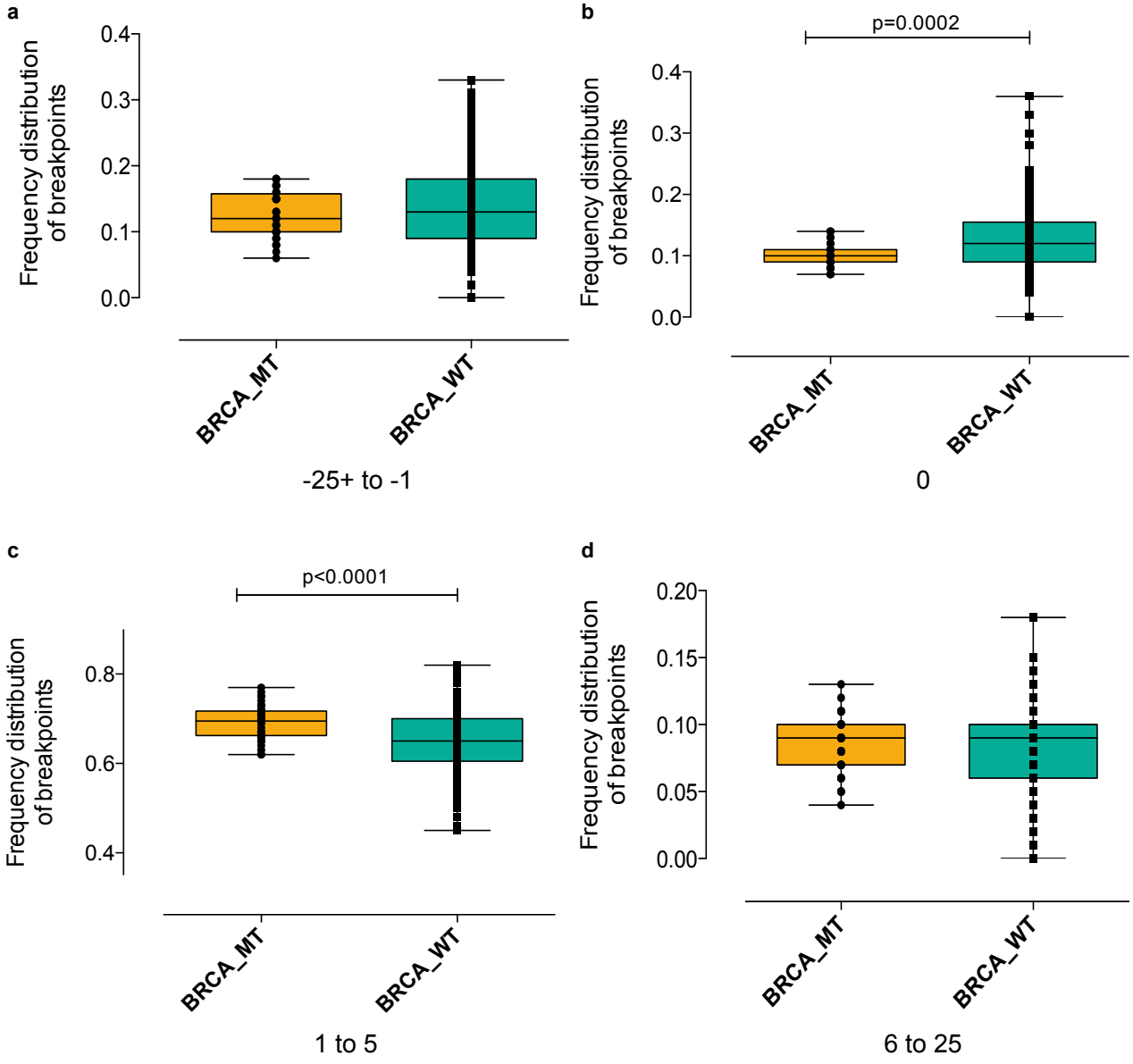


Figure 3-16 Frequency distribution of breakpoints of *BRCA* gene mutation across PDAC, AOCS and OESO samples. *BRCA* mutation status was plotted against the

frequency of somatic rearrangements (y-axis). The sample size for each *BRCA* mutation status was as follows: BRCA_MT (n=40) and BRCA_WT (n=182). Boxplots were coloured according to *BRCA* mutation status – tumours with *BRCA* mutant (BRCA_MT), orange; tumours with *BRCA* wild type (BRCA_WT), green. (a) The categorized length of homology -25+ to -1 bp denotes the characteristic of free DNA insertion at the break. (b) The categorized length of homology 0 bp denotes the characteristic of blunt end. (c) and (d) The categorized length of homology 1 to 5 bp and 6 to 25 bp denote the characteristic of microhomology. All p-values are from Wilcoxon rank-sum test. Comparisons with p-value >0.05 were not labelled.

Although both are members of the HR pathway, studies have shown that *BRCA*1 and *BRCA*2 have distinct function in DNA repair. Therefore, mutations in *BRCA*1 and *BRCA*2 may result in different effects in cancers (Gudmundsdottir and Ashworth, 2006; Liu et al., 2012; Liu and West, 2002). To evaluate if there is difference in breakpoints characteristics of tumours with *BRCA*1 or *BRCA*2 gene mutations, 19 tumours with *BRCA*1 mutation (1 PDAC and 18 AOCS) and 9 tumours with *BRCA*2 mutation (5 PDAC, 3 AOCS and 1 OESO) were investigated.

Interestingly, when comparing the breakpoint characteristics of somatic rearrangements, the result showed that the frequency distribution of microhomology and non-templated sequence insertion characteristics were different between tumours with *BRCA*1 and *BRCA*2 mutation (p=0.0123 and p=0.0083 respectively, Wilcoxon rank-sum test) (Figure 3-17).

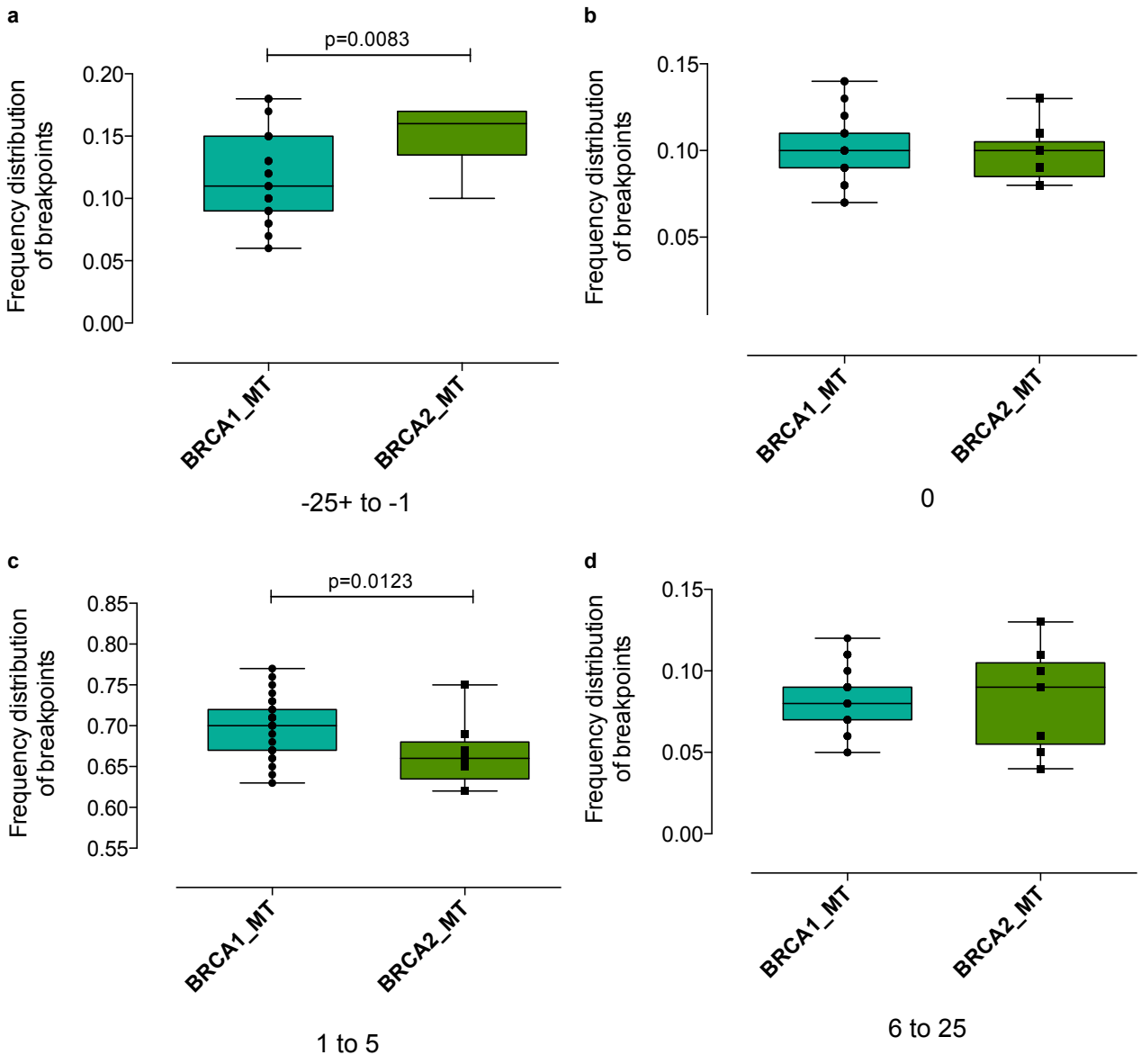


Figure 3-17 Frequency distribution of breakpoints between *BRCA*1 and *BRCA*2 mutation samples across PDAC, AOCS, and OESO samples. *BRCA* mutation status was plotted against the frequency of somatic rearrangements (y-axis). The sample size for each mutation status was as followed: BRCA1_MT (n=19) and BRCA2_WT (n=9). Boxplots were coloured according to *BRCA* gene mutation status – samples with *BRCA*1 mutant (BRCA1_MT), blue-green; samples with *BRCA*2 mutant (BRCA2_MT), dark green. (a) The categorized length of homology -25+ to -1 bp denotes the characteristic of free DNA insertion at the break. (b) The categorized length of homology 0 bp denotes the characteristic of blunt end. (c) and (d) The categorized length of homology 1 to 5 bp and 6 to 25 bp denote the characteristic of microhomology. All p-values are from Wilcoxon rank-sum test. Comparisons with p-value >0.05 were not labelled.

Taken together, the results from different data sets show that the frequency of breakpoints with microhomology of 1 to 5 bp was higher and blunt end (0 bp) was lower in tumours with either *BRCA* gene mutation or a high BRCA mutational signature in PDAC and AOCS cancer genomes (Table 3-2). These results suggest that 'error-prone' NHEJ could be an alternative DNA repair mechanism of the somatic rearrangements due to the deficiency of *BRCA* in HR DNA repair mechanism (Patel et al., 2011). Whilst, the frequency of breakpoints with a blunt end was higher in tumours with *BRCA* wild type or a low BRCA mutational signature which may indicate that the 'accurate' NHEJ shares elements with HR pathway (Pfeiffer et al., 2000).

Table 3-2 Summary results of BRCA mutational signature and *BRCA* mutation

		Median of the frequency changes across the categorized length of homology			
	Difference between…	**-25 to -1 bp**	**0 bp**	**1 to 5 bp**	**6 to 25 bp**
PDAC (n=120)	BRCA mutational signature high vs. low	0.155 vs. 0.115 (p = 0.8077)	0.09 vs. 0.14 (p = 0.0137)	0.68 vs. 0.63 (p = 0.0487)	0.085 vs. 0.08 (p = 0.8165)
	BRCA mutation BRCA_MT vs. BRCA_WT	0.16 vs. 0.12 (p = 0.2306)	0.09 vs. 0.13 (p = 0.0189)	0.67 vs. 0.66 (p = 0.4697)	0.09 vs. 0.08 (p = 0.8669)
AOCS (n=80)	BRCA mutational signature high vs. low	0.125 vs. 0.13 (p = 0.3028)	0.1 vs. 0.11 (p = 0.2009)	0.685 vs. 0.66 (p = 0.0463)	0.09 vs. 0.09 (p = 0.5568)
	BRCA mutation BRCA_MT vs. BRCA_WT	0.11 vs. 0.13 (p = 0.0826)	0.1 vs. 0.12 (p = 0.0147)	0.7 vs. 0.65 (p = 0.0031)	0.085 vs. 0.09 (p = 0.6149)
PDAC + AOCS + OESO (n=222)	BRCA mutation BRCA_MT vs. BRCA_WT	0.12 vs. 0.14 (p = 0.1394)	0.1 vs. 0.12 (p = 0.0002)	0.695 vs. 0.65 (p <0.0001)	0.09 vs. 0.09 (p = 0.8442)
	BRCA mutation BRCA1_MT vs. BRCA2_MT	0.11 vs. 0.16 (p = 0.0083)	0.1 vs. 0.1 (p = 0.7503)	0.7 vs. 0.6 (p = 0.0123)	0.08 vs. 0.09 (p = 0.6566)

3.3.5 Pattern of DNA sequence surrounding the breakpoints

There are many ways a cell acquires DNA breaks. Here, instead of genomic sequence at the breakpoints, the sequences surrounding the breaks were analysed. Studies have suggested that regions of the genomes such as fragile sites and repeats are prone to chromosomal breakage and these regions display certain DNA features (Fungtammasan et al., 2012; Lee et al., 2012; Prak and Kazazian, 2000). In this section, 200 bp regions flanking each side of the breakpoints were used as input to search for DNA enrich motifs using MEME Suite motif finder (Bailey et al., 2009).

3.3.5.1 Pattern of DNA sequence among the genome subtypes of PDAC

A total of 19,482 breakpoints were examined – 4,242 breakpoints from focal subtype (n=32), 9,474 from scattered (n=64), 714 from stable (n=15) and 5,052 from unstable (n=9).

A total of 92 short motifs were discovered across the four genome subtypes of PDAC (Figure 3-18). The top five most frequent motifs in each subtype were manual curated for their contribution in other subtypes. Only the motif 'RAAATA' was shared between focal and stable subtypes at a frequency of 30% and 34% respectively (Table 3-3). In this case, the consensus DNA pattern of motif 'RAAATA' is 'AAAATA' with nucleotide degeneracy at the 1st position. However, other motifs with A+Ts enrichment were also observed in other subtypes (such as 'ARAGAAA', 'RAAATR', 'AWATAY', 'AAATRY', 'AAANAAAA', 'AAATAAW'). This signal of A enrichment was previous reported to be involved in chromosomal fragile sites or TE integration sites and showed to generate cancer-specific rearrangements (such as translocations and deletions) in human cancers (Durkin et al., 2008; Gandhi et al., 2010; Glover and Stein, 1988; Lee et al., 2012).

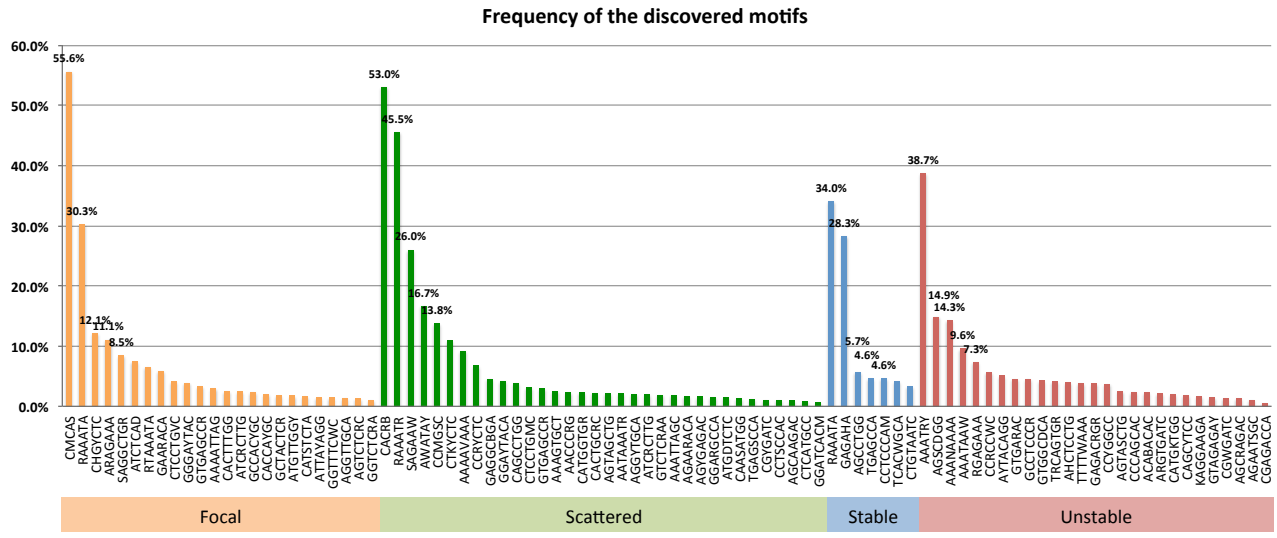


Figure 3-18 Consensus motifs at breakpoint junction in each genome subtype. The frequency of the enriched motifs was plotted against the 92 discovered motifs. The graph is ordered by the frequency of matching motif patterns in each subtype. Bars are coloured according to the genome subtype – focal, orange; scattered, green; stable, blue; unstable, red. The abbreviation for the degenerate bases is found in Supplementary Table 3-2.

Table 3-3 Comparison of the discovered motifs across genome subtypes of PDAC

Motifs	Focal	Stable	Scattered	Unstable
CMCAS	55.6%	-	-	-
RAAATA	30.3%	34.0%	-	-
CHGYCTC	12.1%	-	-	-
ARAGAAA	11.1%	-	-	-
SAGGCTGR	8.5%	-	-	-
GAGAHA	-	28.3%	-	-
AGCCTGG	-	5.7%	-	-
TGAGCCA	-	4.6%	-	-
CCTCCCAM	-	4.6%	-	-
CACRB	-	-	53.0%	-
RAAATR	-	-	45.5%	-
SAGAAW	-	-	26.0%	-
AWATAY	-	-	16.7%	-
CCMGSC	-	-	13.8%	-
AAATRY	-	-	-	38.7%
AGSCDGG	-	-	-	14.9%
AAANAAAA	-	-	-	14.3%
AAATAAW	-	-	-	9.6%
RGAGAAA	-	-	-	7.3%

3.3.5.2 Pattern of DNA sequence surrounding the breakpoints of event types

The involvement of chromosomal fragile sites or TE integration sites has been shown to generate cancer-specific rearrangement (such as translocations and deletions) in human cancers (Dillon et al., 2010; Gandhi et al., 2010; Lee et al., 2012; Popescu, 2003). Here, I conducted analysis of DNA motifs to evaluate if there is a preferred DNA sequences at the surrounding regions of the breakpoints across various event types (i.e. deletions; inversions - comprised of inversions, foldback inversions and amplified inversions; duplications - comprised of duplications and tandem duplications; intra-chromosomal rearrangements; translocations).

A total of 19,482 sequences surrounding the breakpoint junctions at either side was examined (200 bp from each side of the breakpoints) and discovered 97 consensus motifs across the various event types: 15 motifs in deletions, 27 in inversions, 7 in duplications, 31 in intra-chromosomal rearrangements and 17 in translocations (Figure 3-19). The manual curation of the top five most frequent motifs in each event type revealed that most of the discovered motifs were unique to the respective event types except ‘RAAATR’,

which were shared between deletion and inversion event types (Table 3-4). This motif 'RAAATR' represents a consensus DNA pattern of 'AAAATA' with some nucleotide degeneracy at the 1st and 6th position.

The signal of As enrichment has been seen across different event types of pancreatic tumours. In particular, the duplication event type revealed a DNA pattern of 'AAAAAAA' with some nucleotides degeneracy at 5th and 6th position (derived from the motif 'AAAAWWA') as the top most frequent DNA pattern in the regions surrounding the breakpoints (Figure 3-20). The enrichment of As observed here is consistent with what have been previously reported whereby the transposon element insertion (TEI) resulted in duplication of target sites and were frequently found in As rich DNA regions (Lee et al., 2012; Prak and Kazazian, 2000). Therefore, the signal of As in duplication event type highlights the possibility of retrotransposition activity presents in this cohort of pancreatic cancer genomes.

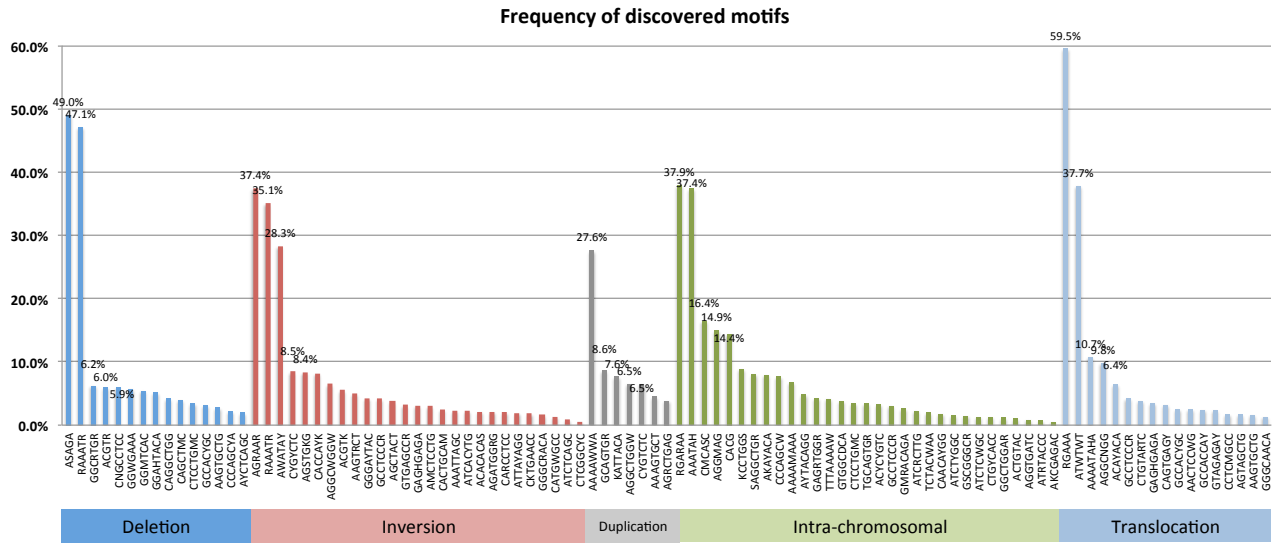


Figure 3-19 Consensus motifs at breakpoint junction in each event type. The frequency of the enriched motifs was plotted against the 92 discovered motifs. The graph is ordered by the frequency of matching motif patterns in each genome subtype. Bars are coloured according to the event type – deletion, blue; inversion, red; duplication, grey, intra-chromosomal, green; translocation, light blue. The abbreviation for the degenerate bases is found in Supplementary Table 3-2.

Table 3-4 Comparison of the discovered motifs across event types of PDAC

Motifs	Deletion	Duplication	Intra-chr	Inversion	Translocation
ASAGA	49.0%	-	-	-	-
RAAATR	47.1%	-	-	35.1%	-
GGCRTGR	6.2%	-	-	-	-
ACGTR	6.0%	-	-	-	-
CNGCCTCC	5.9%	-	-	-	-
AAAAWWA	-	27.6%	-	-	-
GCAGTGR	-	8.6%	-	-	-
KATTACA	-	7.6%	-	-	-
AGGCTGGW	-	6.5%	-	-	-
CYGTCTC	-	6.5%	-	-	-
RGARAA	-	-	37.9%	-	-
AAATAH	-	-	37.4%	-	-
CMCASC	-	-	16.4%	-	-
AGGMAG	-	-	14.9%	-	-
CACG	-	-	14.4%	-	-
AGRAAR	-	-	-	37.4%	-
AWATAY	-	-	-	28.3%	-
CYGYCTC	-	-	-	8.5%	-
AGSTGKG	-	-	-	8.4%	-
RGAAA	-	-	-	-	59.5%
ATWTWT	-	-	-	-	37.7%
AAAATAHA	-	-	-	-	10.7%
AGGCNGG	-	-	-	-	9.8%
ACAYACA	-	-	-	-	6.4%

Duplication

166/602
(27.6%)

Figure 3-20 Consensus motif in the surrounding sequence of duplications. The enriched motifs were generated from 166 sequences that surround the breakpoint junction at either side of duplication event type. The consensus pattern of duplication is 'AAAAAAA' with some nucleotides degeneracy at 5th and 6th position.

3.4 Discussion

In this chapter, I presented an analysis of somatic rearrangement breakpoints in pancreatic primary tumours. Mapping the breakpoints to base resolution allows precise annotation of the chromosomal rearrangements and the identification of characteristics of breakpoints could give us clues about the potential mechanisms involved in DNA repair and rearrangements formation.

The compendium of somatic rearrangements obtained from 120 pancreatic tumours showed that each pancreatic cancer genome displayed distinct patterns of somatic rearrangements and breakpoints characteristics. A total of 10,074 high confidence somatic rearrangements were identified across these tumours. Intra-chromosomal rearrangements were the most prevalent somatic event type and inversion formed the most common subcategory. Of 10,074 somatic rearrangements, 95% of the somatic rearrangements were identified with split contig alignment and could be used to characterise the breakpoints junctions. The majority of the breaks exhibit the microhomology characteristic (74.8%). Approximately 67% of these somatic rearrangements harboured 1 to 5 bp short overlapping bases suggesting that the NHEJ might be the predominant DNA repair mechanisms in this cohort of pancreatic cancer genomes (Chen, 2001; Yang et al., 2013). However, it should also be noted that the different length of homology are shared across multiple mechanisms according to these literatures (Conrad et al., 2010; Hastings et al., 2009; Lam et al., 2010; Lee et al., 2007; Yang et al., 2013). For example, NHEJ is able to generate deletions with blunt end (0 bp), microhomology (length 1 to 5 bp), non-templated sequence insertion (length 1 to >20 bp) and genomics shards (length >20 bp).

Recent work by the members of QCMG showed that the genomic distribution of somatic rearrangements in cancer genomes could stratify pancreatic cancer into 4 genome subtypes – focal, scattered, stable and unstable (Waddell et al., 2015). The unstable subtype was frequently associated with *BRCA* mutation and a high BRCA mutational signature. Tumours with unstable genomes responded better to platinum-based therapy suggesting that unstable genome could be potential candidate biomarker for therapeutic responsiveness. However, in that study approximately 28% of tumours with unstable genome did not harbour *BRCA* mutation or had low BRCA mutational signature indicating that there are other mechanisms resulting in genome instability. Here, we looked into the characterization of breakpoints junctions and the sequence context surrounding

the breakpoints across a large cohort of primary pancreatic tumours to investigate what are the potential mechanisms driving the formation of somatic rearrangements.

BRCA genes are known to be involved in DNA repair via the HR pathway (Gudmundsdottir and Ashworth, 2006; Zhang and Powell, 2005) while the characteristic of microhomology has been associated with NHEJ repair mechanism (Chen, 2001; Yang et al., 2013). The results presented here showed that *BRCA* deficient tumours (point mutation and high BRCA mutational signature) of both PDAC and AOCS often have higher frequency of somatic breakpoints with microhomology of 1 to 5 bp. This suggests that due to the deficiency of *BRCA* in the HR pathway, the DNA repair machinery of the cancer cells might switch to the 'error-prone' NHEJ, which uses microhomology as an alternative mechanism in order to repair the broken chromosomes (Patel et al., 2011). Otherwise, for cases where HR mechanism is still intact, the 'accurate' NHEJ could re-join the chromosomal breakage accurately and generate somatic rearrangements with blunt end characteristic (Pfeiffer et al., 2000). In agreement with the genome subtype analysis whereby the unstable subtype that was associated with HR defective pathway, has a lower frequency of breakpoints of blunt end characteristic when comparing to the other subtypes. In addition, there may be a difference in breakpoint characteristics pattern between *BRCA*1 and *BRCA*2 suggesting that they might have a distinct function in the HR pathway.

*BRCA*1/2 genes play an important role in DNA repair mechanisms. It has been shown that tumours with defective BRCA pathway responded better to platinum-based therapy (D'Andrea, 2003). Cells with *BRCA1* or *BRCA2* mutation are highly sensitive to PARP1 inhibition. The loss of PARP1 activity seems to lead to an accumulation of multiple DNA lesions and result with an increase chromosome instability leading specifically tumour cell death (Farmer et al., 2005). As previous work from QCMG showed that a *BRCA* mutation does not always correlate with a BRCA-like mutational signature suggesting that some mutations are not significant, while a tumour can have a BRCA-like mutational signature but no *BRCA1/2* mutation may suggest that other members of HR pathway are mutated or perturbed. Therefore, the different breakpoint characteristics between tumours with *BRCA* mutation and BRCA mutational signature reinforce previous evidence that a subtype of pancreatic tumour might have deficiency in the HR pathway and could response to PARP1 inhibitors.

One important weakness in these analyses is the inability of detecting extensive homology at the chromosomal breakage leading to a bias observation toward microhomology at breakpoints. However, this is a limitation of most structural variation detection tools using short sequencing reads (≤300 bp) (Onishi-Seebacher and Korbel, 2011) but could possibly be resolved by using the alternative sequencing technologies that produces long reads of up to 15 kb.

The analysis of DNA sequences surrounding the breakpoints has provided insight of biological activities or mechanisms driving the formation of somatic rearrangements. We learnt that these formations of somatic rearrangements in pancreatic tumours among events such as genome subtype and event type have a significant preference for DNA regions enriched for A+Ts. Possible mechanisms that could associate with these observations are:

(1) Retrotransposition activity. Somatic transposon element (TE) insertions have been shown to be involved in tumorigenesis in other human cancers (Helman et al., 2014; Lee et al., 2012). Our results showed that majority of the somatic rearrangements displayed enrichment of A+Ts around the DNA sequence surrounding the breakpoints throughout the four genome subtypes and different event types. The motifs in genome subtypes are ‘RAAATA’, ‘ARAGAAA’, ‘RAAATR’, ‘AWATAY’, ‘AAATRY’, ‘AAANAAAA’, and ‘AAATAAW’. For the different event types, the motifs include ‘RAAATR’, ‘AAAAWWA’, ‘AAATAH’, ‘AWATAY’, ‘ATWTWT’, and ‘AAAATAHA’. In particular, the most frequent DNA motif in duplication event was AAAAWWA represented by a consensus DNA pattern of ‘AAAAAAA’. Previous studies have reported similar consensus motif patterns (e.g. ‘TTAAAA’ and ‘TTTTAAAAA’) at the target site of somatic retrotransposons insertions associated with duplications of human cancer genomes (Helman et al., 2014; Lee et al., 2012; Prak and Kazazian, 2000).

(2) Association with chromosomal fragile sites. The fragile sites have been described to be associated with A+Ts rich regions and also known as one of the contributors to genomic instability in human disease especially cancers (Arlt et al., 2006; Barlow et al., 2013; Gandhi et al., 2010). Studies have shown that fragile sites could arise from chromatin modification (Jiang et al., 2009) and hypomethylation (Shann et al., 2008) which play a role in the formation of structural rearrangements in human cancers.

In summary, the different composition of event types and differences between the breakpoints characteristics across the pancreatic primary tumours suggest that the underlying mechanisms of somatic rearrangements formation in pancreatic carcinogenesis are complex with potentially multiple mechanisms acting within a single genome. Overall, the analyses highlight that NHEJ is an active DNA repair mechanism during the development of pancreatic cancer. The identification of breakpoints with microhomology may indicate that HR pathway is defective in a subgroup of pancreatic tumours. Furthermore, the analysis of the DNA sequences surrounding the breakpoints showed the enrichment of A+Ts that might indicate that retrotransposons and fragile sites could associate with the number and the distribution of somatic rearrangements in pancreatic cancer, however, this aspect of the analysis need to be further investigated.

3.5 Supplementary Material

Supplementary Table 3-1 Summary of verified somatic rearrangements breakpoints

Donor id	Event type	chr from	chr from breakpoint	chr from strand	chr to	chr to breakpoint	chr to strand	Verified	Identified breakpoints	Strand info
APGI_1959	Intra-chromosomal	10	127115886	+	10	115575262	+	no	-	
APGI_1959	Intra-chromosomal	10	128138324	+	10	132334501	+	no	-	
APGI_1959	Inversion	10	128143345	-	10	131927831	+	no	-	
APGI_1959	Intra-chromosomal	10	131182809	+	10	124275387	+	no	-	
APGI_1959	Intra-chromosomal	10	131357885	+	10	80347920	+	no	-	
APGI_1959	Inversion	10	81884712	-	10	131680198	+	yes	-	
APGI_1959	Intra-chromosomal	10	81934545	+	10	81937501	+	yes	-	
APGI_1959	Deletion	10	112178601	+	10	112224869	+	yes	-	
APGI_1959	Amplified Inversion	10	131220180	-	10	131223531	+	yes	-	
APGI_1959	Intra-chromosomal	10	131679280	+	10	81934364	+	yes	-	
APGI_1959	Deletion	10	132520931	+	10	132626399	+	yes	-	
APGI_1959	Amplified Inversion	10	132691801	+	10	132692644	-	yes	-	
APGI_1959	Intra-chromosomal	10	127115886	+	10	115575262	+	no	-	
APGI_1959	Deletion	10	90698204	+	10	118965300	+	yes	chr10: 118965300-90698204	-/-
APGI_1959	Deletion	10	111236293	+	10	125812179	+	yes	chr10:111236293-125812179	+/+
APGI_1959	Intra-chromosomal	10	125985379	+	10	111257993	+	yes	chr10:111236293-125812179	+/+
APGI_1959	Foldback Inversion	10	111822448	+	10	116136459	-	yes	chr10:116136459-111822448	+/-
APGI_1959	Foldback Inversion	10	112783819	+	10	117872128	-	yes	chr10:117872128-112783819	+/-
APGI_1959	Foldback Inversion	10	117323037	-	10	118948105	+	yes	chr10:118948105-117323037	-/+
APGI_1959	Intra-chromosomal	10	124171213	+	10	110529485	+	yes	chr10:124171213-110529485	+/+
APGI_1959	Intra-chromosomal	10	126565271	+	10	126918647	+	yes	chr10:126565264-126918647	+/+
APGI_1959	Intra-chromosomal	10	127060916	+	10	125985476	+	yes	chr10:127060916-125985476	+/+
APGI_1959	Amplified Inversion	10	130765200	-	10	130765864	+	yes	chr10:130765200-130765864	-/+
APGI_1959	Tandem	10	131183502	+	10	127081152	+	yes	chr10:131183502-	+/+

	Duplication								127081152	
APGI_1959	Duplication	10	131553675	+	10	104959011	+	yes	chr10:131553675-104959011	+/+
APGI_2049	Intra-chromosomal	7	145759290	+	7	109067344	+	no	-	
APGI_2049	Foldback Inversion	7	146696184	-	7	147399313	+	no	-	
APGI_2049	Intra-chromosomal	6	162287635	+	6	162295523	+	no	-	
APGI_2049	Foldback Inversion	3	20319547	+	3	174852757	-	yes	-	
APGI_2049	Translocation	7	109066841	+	3	191369830	+	yes	-	
APGI_2049	Translocation	7	109613754	+	3	21654066	+	yes	-	
APGI_2049	Translocation	7	114376753	+	3	175416919	+	yes	-	
APGI_2049	Intra-chromosomal	9	9542252	+	9	9544210	+	yes	-	
APGI_2049	Translocation	15	29104579	+	11	31548728	+	no	-	
APGI_2049	Inversion	18	18530073	+	18	19381954	-	no	-	
APGI_2049	Intra-chromosomal	13	68878214	+	13	69058982	+	yes	-	
APGI_2049	Deletion	17	11097052	+	17	12862481	+	yes	chr17:12862481-11097058	-/-
APGI_2049	Amplified Inversion	17	15210082	-	17	15211382	+	yes	chr17:15210082-15211382	-/-
APGI_2049	Inversion	17	15281828	+	17	50164473	-	yes	chr17:50164473-15281828	+/-
APGI_2049	Inversion	18	18777462	-	18	19385721	+	yes	chr18:19385724-18777462	-/+
APGI_2049	Deletion	20	61035468	+	20	61322801	+	yes	chr20:61035473-61322805	+/+
APGI_2049	Intra-chromosomal	23	7245498	+	23	7190442	+	yes	chr23:7190442-7245504	-/-
APGI_2049	Translocation	3	191163264	+	7	146440446	+	yes	chr7:146440446-chr3:191163270	-/-
APGI_2049	Translocation	3	174865868	+	7	146704618	-	yes	chr7:146704618-chr3:174865879	+/-
APGI_2049	Foldback Inversion	7	89045573	+	7	107390164	-	yes	chr7:89045585-107390164	+/-
APGI_2156	Intra-chromosomal	2	86639146	+	2	86640479	+	no	-	
APGI_2156	Deletion	16	78661636	+	16	78663235	+	no	-	
APGI_2156	Intra-chromosomal	22	36834501	+	22	36840002	+	no	-	
APGI_2156	Foldback Inversion	4	184211805	-	4	184545554	+	yes	-	
APGI_2156	Intra-chromosomal	6	44601759	+	6	44601823	+	no	-	
APGI_2156	Intra-chromosomal	7	48898896	+	7	48899988	+	no	-	
APGI_2156	Deletion	12	52330742	+	12	52426239	+	no	-	
APGI_2156	Foldback Inversion	1	228189395	+	1	228316922	-	yes	chr1:228189395-228316922	+/-
APGI_2156	Foldback Inversion	1	228285225	-	1	228324721	+	yes	chr1:228325035-228285452	+/-

APGI_2156	Deletion	10	84254184	+	10	84287402	+	yes	chr10:84254182-84287402	+/+
APGI_2156	Deletion	16	78617053	+	16	78644950	+	yes	chr16:78617053-78644951	+/+
APGI_2156	Deletion	16	78715249	+	16	78782389	+	yes	chr16:78715249-78782389	+/+
APGI_2156	Tandem Duplication	20	14625647	+	20	14574229	+	yes	chr20:14625651-14574229	+/+
APGI_2156	Deletion	3	60499811	+	3	60615358	+	yes	chr3: 60499811-60615358	+/+
APGI_2156	Deletion	3	116111888	+	3	116755875	+	yes	chr3:116755875-116111887	-/-
APGI_2156	Intra-chromosomal	3	60345191	+	3	60339070	+	yes	chr3:60345191-60339072	+/+
APGI_2156	Foldback Inversion	4	184211274	+	4	184567325	-	yes	chr4:184567323-184211274	+/-
APGI_2156	Deletion	4	184271711	+	4	184571354	+	yes	chr4:1845713454-184271711	-/-
APGI_2156	Intra-chromosomal	9	21985014	+	9	22621446	+	yes	chr9: 21985014-22621446	+/-

Supplementary Table 3-2 Abbreviations for degenerate bases used in this chapter

Degenerate base	Actual bases
N	A or C or G or T
B	C or G or T
D	A or G or T
H	A or C or T
V	A or C or G
K	G or T
M	A or C
R	A or G
S	C or G
W	A or T
Y	C or T

4 Chapter 4

Identification of personalized DNA-based biomarkers for pancreatic cancer and the assessment of whether they can be used to monitor tumour burden and response to chemotherapy

4.1 Introduction – Somatic mutations as biomarkers in circulating cell free DNA

The definition of a cancer biomarker by National Cancer Institute is “a biological molecule found in blood, other body fluids or tissues that is a sign of normal or abnormal process, or of a condition or disease”. Somatic mutations such as structural rearrangements or point mutations are features of cancer genomes. Thus, they hold great promise for patient specific biomarkers which can be used to either predict response to treatment or monitor relapse. Previous studies have shown that the primary tumour releases DNA fragments into the circulation carrying characteristics of the tumour, which could reflect diverse pathological states (Leary et al., 2010; McBride et al., 2010).

4.1.1 Approaches for monitoring circulating tumour DNA

Over the years, studies have demonstrated the quantification of tumour specific mutations in circulating cell free DNA (cfDNA) using digital PCR (Sykes et al., 1992; Vogelstein and Kinzler, 1999) or next generation sequencing (Rusk and Kiermer, 2008).

Studies have shown that tumour specific chromosomal rearrangements are detectable in circulating cfDNA obtained from serum or plasma of cancer patients. Leary et al. (Leary et al., 2010) developed an approach called “Personalised Analysis of Rearrangement Ends” (PARE) to identify somatic rearrangements that are unique to each patient. This approach uses next generation sequencing of the primary tumour to detect tumour specific somatic rearrangements; a quantitative PCR assay is then used to detect and quantify the rearrangements in the cfDNA. McBride et al. presented a similar assay to detect and quantify tumour specific rearrangements in plasma or serum of patients with metastatic breast and bone cancers (McBride et al., 2010).

Another approach is to detect specific somatic mutations in cfDNA, which are at higher frequency in a particular cancer. These mutations could be “hotspot” mutations, which occur at the same position within a gene. For example, *PIK3CA* in breast cancer

(Beaver et al., 2014), *KRAS* and *BRAF* in colorectal cancer and *EGFR* in non-small cell lung cancers (Yung et al., 2009). The digital PCR (dPCR) technology is used and directly counts the number of target molecules to give an absolute quantification of any rare allele without relying on reference standards or endogenous controls. PCR reaction is partitioned into hundreds or thousands individual PCR reactions for amplification depending on the platform. The number of amplified molecules will be estimated by Poisson algorithm. A number of vendors have launched commercial instruments that allow this technology to be used for quantification of ctDNA (Table 4-1). Of these, two dPCR instruments were evaluated in this chapter to quantify circulating cfDNA. A detailed review of dPCR instruments has been reported by Monya Baker (Baker, 2012).

Table 4-1 Comparison of potential dPCR instruments for circulating DNA project

Name of the instruments (Vendor)	Chemistry	Sample per plate/chip/strip	Number of reactions per sample	Reaction volume
Quantstudio® 3D digital PCR System (Life Technologies)	chip-based	48 samples/plate	64 reactions	100 µL
BioMark™ HD System (Fluidigm Corporation)	chip-based	12 samples/chip	765 reactions	8 µL
QX200 Droplet Digital PCR (Bio-Rad)	droplet-based	8 samples/strip	20,000 droplets*	20 µL
RainDrop™ Digital PCR System (RainDance Technologies)	droplet-based	8 samples/strip	10,000,000 droplets*	up 50 µL

*A droplet is equivalent to one reaction

With the advent of genomic technologies, next generation sequencing technology is an alternative approach for mutation detection in cfDNA. Different sequencing methods have been used to detect tumour specific mutations in plasma or serum of cancer patients (Leary et al., 2012; Murtaza et al., 2013; Narayan et al., 2012). This includes whole genome sequencing, exome sequencing, and target amplicon sequencing. Whole genome sequencing and exome sequencing can identified novel mutations without the needs to focus on reported mutations. In contrast, targeted amplicon sequencing focuses on a single or panel of predefined or hotspot mutations. In principle, the quantification using sequencing technologies make use of the knowledge of genomic position of known mutations which the frequency can be read out directly from the sequencing data.

However, the minimum mutant allele frequency that can be detected, will depend on the read depth background rates of non-reference reads (error rate).

4.1.2 Current biomarker for pancreatic cancer

The most widely used and investigated biomarker for pancreatic cancer is cancer antigen 19.9 (CA 19.9). It was first discovered through monoclonal antibody studies of colorectal cancer cell lines (Koprowski et al., 1979), and was later showed that pancreatic cancer patients have increased levels of CA 19.9 (Steinberg, 1990). However, CA 19.9 marker has limitations such as inadequate sensitivity and specificity. Elevated levels of CA 19.9 are observed in other gastrointestinal cancers including gastric and colorectal cancers (Duffy, 1998). CA 19.9 lacks sensitivity in detecting small pancreatic tumours (Steinberg, 1990). Assays for CA 19.9 have been inconsistent producing different results in different laboratories (Duffy et al., 2010) and the expression of CA 19.9 is only subjected to population who carries Lewis A blood group antigen (Takasaki et al., 1988). Together, these limitations reinforce the need for a more accurate biomarker in clinical management for pancreatic cancer patients.

4.1.3 Alternative biomarkers in pancreatic cancer

Cancer genomes harbour tens to thousands of somatic mutations in its pathogenesis and could be used as potential biomarkers (Birnbaum et al., 2011; Campbell et al., 2010). The PARE assay requires sequencing of a primary tumour (Leary et al., 2010), however, in the case of pancreatic cancer, tumour material is not available for 80% of patients whom have non-resectable disease. However, it is known that the mutation of *KRAS* gene is found in over 90% of pancreatic cancer (Thomas et al., 2007) and most of these mutations occur at G12 amino acid.

Therefore, in this chapter I explore both PARE assay (personalized approach) and the quantification of *KRAS* mutation (generic approach) in circulating cfDNA with the aim to investigate whether such mutations could be used as biomarkers in pancreatic cancer.

4.2 Material and methods

Different approaches were used to test the use of tumour specific somatic mutations as ctDNA biomarkers to monitor tumour burden and response to chemotherapy in patients with pancreatic cancer. A schematic of the work is illustrated in Figure 4-1. Essentially, two approaches were used: personalized tumour specific rearrangements and generic recurrent *KRAS* mutation as biomarkers. I first designed primers and probes targeting small amplicons (<200 bp) which spanned the rearrangement breakpoints or *KRAS* mutation. I then attempted to quantify the mutations in serum or plasma obtained from patients with pancreatic cancer during the disease course to evaluate the association of mutation load in circulation with tumour burden of individuals.

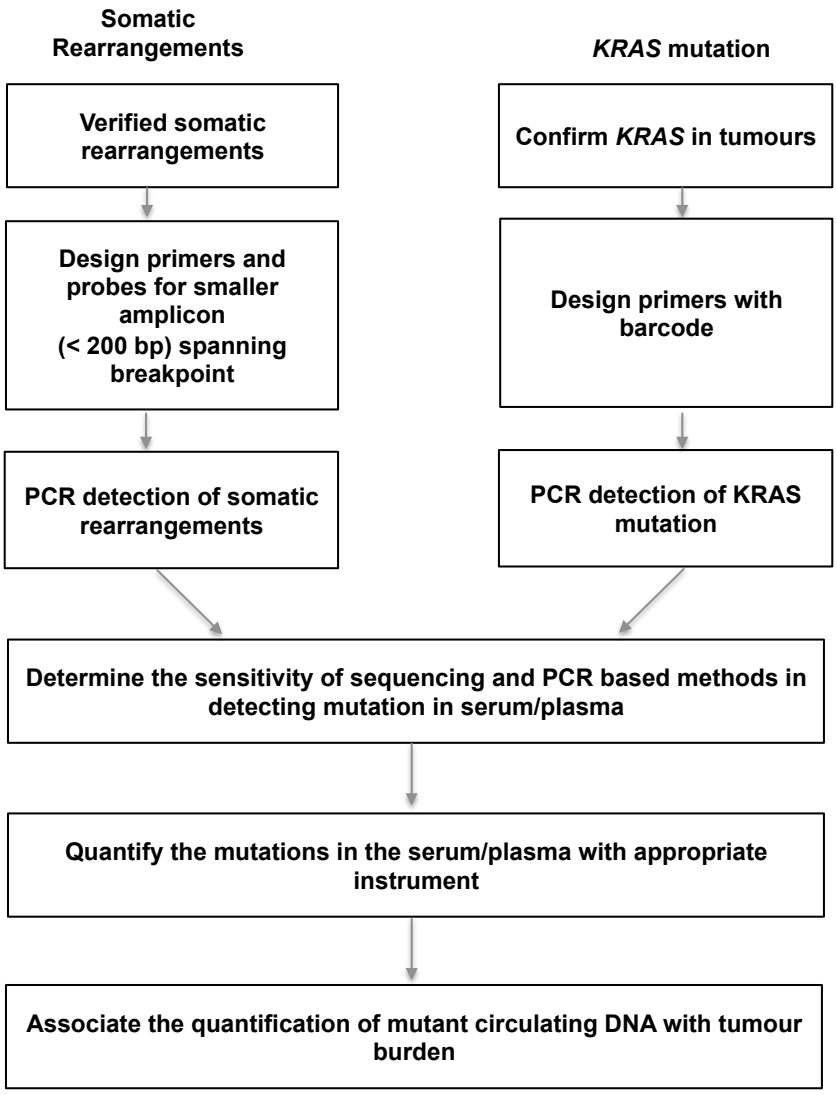


Figure 4-1. Schematic analysis for detecting tumour specific mutations in serum or plasma.

4.2.1 Patient serum or plasma samples

Patients were recruited to the Australian Pancreatic Cancer Genome Initiative (APGI) and have given informed written consent for the participation. A total of 28 serum or plasma samples were collected from five patients (APGI 1830, 1953, 1959, 2302, 2353) during their course of therapy. Circulating DNA was isolated from 0.8 mL to 1.7 mL serum or plasma using QIAamp Circulating Nucleic Acid kit (Qiagen) following the manufacturer's instruction. ***Note: Patients were recruited by the APGI Team coordinated by Amber Johns, and the extraction of DNA from the serum was performed by Dr. Venessa Chin, Garvan Institute of Medical Research.***

4.2.2 Quality control and quantification of serum or plasma DNA

Upon receiving the serum or plasma DNA samples, the size, quality and quantity of the serum or plasma DNA were assessed on Agilent 2100 BioAnalyzer using the High Sensitivity DNA Kit. An example of Bioanalyzer profile of sample that passed and failed the quality check is shown in Figure 4-2.

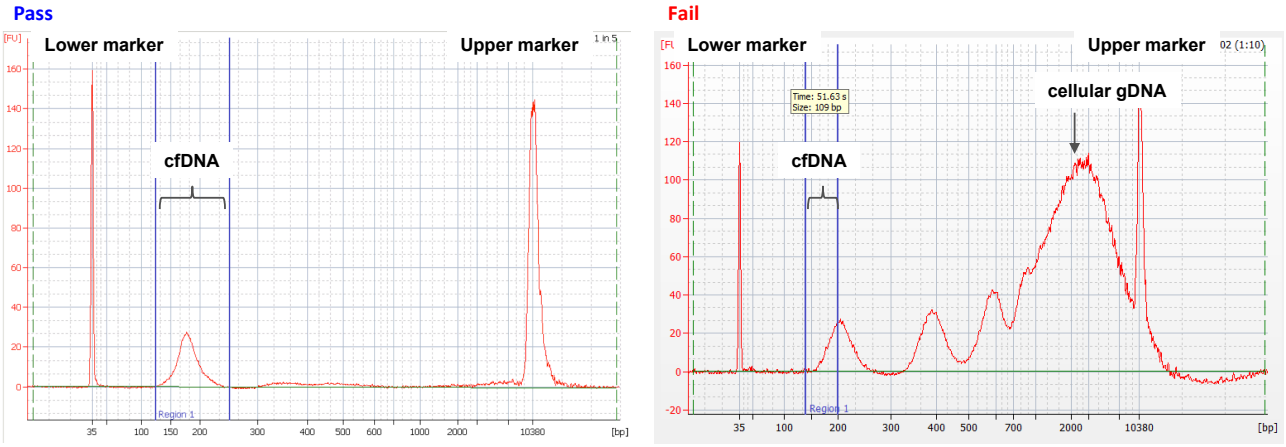


Figure 4-2 Agilent profile showing pass and fail of the QC check for cfDNA. The size of the nucleic acid is on the x-axis and the fluorescent intensity is on y-axis. The vertical blue indicate the expected size of cfDNA (approximately 130 to 200 bp). Sample that passes QC indicates that the size of the serum or plasma DNA lies within expected range. For a sample that fails QC, the DNA profile indicates that cfDNA is contaminated with genomic DNA.

Of the 28 samples received, 15 serum or plasma DNA were detected within the expected size between 130 and 200 bp, however, many were in small quantity and volume (0.14 ng/uL to 5.3 ng/uL; >10 uL to 20 uL) (Supplementary Table 4-1). The remaining 13

did not pass the quality check as they were either contaminated with genomic DNA or DNA was below the detection limits.

4.2.3 Whole genome amplification of serum or plasma DNA

Illumina Genomic DNA Sample Prep Kit was used for whole genome amplification (WGA) of the circulating cfDNA. The amplification was carried out by the ligation of adaptors to cfDNA (fragments 130 to 200 bp) and PCR amplification (i.e. 18 cycles), following the manufacturer's instruction. The amount of serum or plasma used for amplification and quantification is listed in Supplementary Table 4-2.

4.2.4 Personalised analysis – Somatic rearrangement

Identification of somatic rearrangements

Whole genome sequencing of primary tumour and matched normal from patient APGI 1959 was performed and 27 somatic rearrangements were identified using qSV which was previously described in Chapter 3. A somatic deletion of 50 kb spanning across chr10:81,937,567-131,680,490 was selected as a potential personalised biomarker. This event resulted a deletion of the *PTEN* gene and so was a suitable event for testing, as it is likely to be important to the tumour, therefore will not get lost during tumour progression.

Primers and probes for somatic rearrangement

Primers and probes were designed to detect tumour specific rearrangement (deletion) in the serum using custom TaqMan® assay design tool (Invitrogen) and generated a PCR product of 180 bp. The TaqMan® assay includes DNA template, a pair of PCR primers and a sequence-specific TaqMan® probes (Supplementary Table 4-3).

4.2.4.1 Digital PCR of somatic rearrangement

BioMark system (Fluidigm)

Digital PCR (dPCR) was carried out in 12.765 digital array chip which consists of 12 panels (sample wells) and each panel partitions the sample premixed with PCR reagents into 765 individual PCR reaction, allowing quantification of target sequences. Each panel was set up with DNA sample, 5 µL TaqMan Gene Expression Master Mix (Applied Biosystems), 0.5 µL of 20X GE sample loading reagent (Fluidigm) and 0.5 µL of 20X TaqMan® gene expression assay in a final volume of 10 µL. Then, the digital array chip was primed and the samples were loaded by NanoFlex 4IFC Controller (Fluidigm). The PCR amplification was carried in the BioMark Real-time PCR system using a cycle profile

of initial at 50°C for 2 min and incubation at 95°C for 10 min followed by 40 cycles of 95°C for 15 sec and 58°C for 1 min. Digital PCR analysis was done using BioMark dPCR analysis software version 3 (Fluidigm).

QX100™ Droplet Digital™ PCR System (Bio-Rad)

The emulsion PCR approach used by the droplet dPCR allows partition of the PCR reaction in a higher number of partitions than chip-based increasing the dynamic range of the quantification. Each reaction mixture contains 12.5 µL of 2X ddPCR master mix (Bio-Rad), 1.25 µL of 40X TaqMan® gene expression assay, and DNA sample in a final volume of 25 µL. The prepared reactions were loaded into individual wells of the disposable droplet generator cartridge with 70 µL droplet generation oil (Bio-Rad), and then into the QX100 droplet generator (Bio-Rad) to generate ~20,000 emulsion droplets. The emulsified droplets were amplified using a cycling profile of 50°C for 2 min and incubation at 95°C for 10 min followed by 40 cycles of 95°C for 15 sec and 58°C for 1 min. The output was analysed with QuantaSoft analysis software 1.3.2.0 (Bio-Rad).

4.2.4.2 Quantification of ctDNA using somatic rearrangement

Patient APGI 1959: A serum sample was collected during chemotherapy and passed QC. The dPCR technologies (Fluidigm and Bio-Rad) were used to quantify tumour specific rearrangement (deletion) in the circulation of this patient. Taking an assumption that whole genome amplified (WGA) serum has sufficient amount of tumour DNA molecules, 1 in 5 serial dilutions were carried in WGA serum DNA ranging from 1:125 to 1:78,125 to estimate the number of copies of tumour DNA.

4.2.5 Generic analysis – recurrent *KRAS* mutation

Preparation of serial dilution of tumour DNA and germline DNA samples

Primary tumour DNA was mixed at known concentration with matched normal DNA to determine the sensitivity of the quantification of sequencing and dPCR instruments as well as acting as controls when quantifying mutation in cfDNA of cancer patients. Primary tumour DNA was diluted in normal DNA as a 1 in 5 dilution until reaching 1:78,125. The input amount of DNA varied between 5 to 50 ng depending on the methods and/or instruments used. The actual amount of tumour DNA in each serial dilution depends on the cellularity of the primary tumour analysed.

4.2.5.1 Next generation sequencing

Primers for KRAS mutation

Primers were designed to detect *KRAS* mutations at *KRAS* exon 2, chr12: 25398284 (C>T – patient APGI 1953) and exon 3, chr12: 25380277-25380278 (GA>TT – patient APGI 2353). Primers sequences are listed in the Supplementary Table 4-3. For nested PCR, the first pair of primers were designed further away from the target region and generated PCR products of size <200 bp. The second pair of primers was designed nearer to the target region to generate PCR products (Supplementary Table 4-3).

Preparation of sequencing libraries and sequencing

Serial dilutions of primary tumour DNA and plasma DNA samples were subjected to an initial round of 18 PCR cycles, followed by a second round of 35 PCR cycles to incorporate sequencing adaptors and barcodes. The barcoded amplicons (DNA libraries) were analysed using Agilent 2100 BioAnalyzer to ensure expected insert size of <200 bp. The PCR products were pooled and purified using AMPure XP beads using a bead:DNA volume ratio of 1.8:1. The library was then quantified using Qubit Flurometer (Invitrogen) and subjected to Ion Torrent PGM (1 x 200 bp run) and Illumina MiSeq (2 x 150 bp run) amplicon sequencing according to manufacturers' instruction. ***Note: Sequencing was performed by QCMG sequencing team.***

Analysis of sequencing data

Sequencing reads were aligned to the reference genome (hg19) using TMAP (Ion Torrent) or BWA (MiSeq). The frequency of the mutated allele (*KRAS*) was measured by the ratio of the number of reads with the mutant allele and the total read counts at the target position. ***Note: Pre-processing of the data was performed by QCMG informatics team.***

4.2.5.2 Digital PCR of *KRAS* mutation

Primers and probes

Primers and probes were designed to detect *KRAS* reference base (VIC) and mutated base (FAM) at *KRAS* exon 2, chr12: 25398284 (C>T, patient APGI 1959) and exon 3, chr12: 25380277- 25380278 (GA>TT, patient APGI 2353). The TaqMan® assay includes DNA template, a pair of PCR primers and two sequence-specific TaqMan® probes (Supplementary Table 4-4).

BioMark system (Fluidigm)

Refer to previous section.

QX100™ Droplet Digital™ PCR System (Bio-Rad)

Refer to previous section.

4.2.5.3 Quantification of ctDNA using *KRAS* mutation

Patient APGI 1953: The quantification of *KRAS* mutation (chr12: 25398284, C>T) in the plasma of this patient was performed using PGM sequencing method. The serial dilutions of tumour DNA were used as standard controls. The range of tumour DNA per reaction varies from 6.9 ng to 0.0004 ng. Nested PCR was performed to amplify *KRAS* mutation in both WGA cfDNA and standard controls. Serial dilutions of the primary tumour and matched normal DNA, and WGA cfDNA from two different time points of the patient disease course (post-operation and during chemotherapy) were included in the sequencing assay.

Patient APGI 2353: The quantification of plasma was carried out using dPCR method. The quantification of *KRAS* mutation was performed on WGA cfDNA collected during pre-operation and post-operation (after 2^{nd} and 7^{th} day). Serial dilutions of tumour and matched normal DNA (total input of 72 ng (Bio-Rad) and 36 ng (Fluidigm) were prepared as standard controls. Mixtures were diluted 5 times until reaching 1:3,125. Each dPCR assay includes serial dilutions of the primary tumour and matched normal DNA, DNA template, PCR primers pair and sequence-specific TaqMan probe to capture GA>TT in *KRAS* mutation.

To increase the sensitivity, WGA cfDNA were then subjected to an extra round of 18 PCR cycles using primers targeting *KRAS* regions prior to the quantification on dPCR. PCR amplified WGA plasmas were quantified in five replicates on two separate dPCR assays using both Fluidigm and Bio-Rad instruments.

4.3 Results

4.3.1 Personalised analysis – somatic rearrangement

Validation and detection of the tumour specific rearrangement biomarker

Whole genome sequencing identified 27 somatic rearrangements in patient APGI 1959. A somatic deletion of 50 Mb which contains *PTEN* and is located at chr10:81,937,567-131,680,490 was selected as tumour specific rearrangement biomarker. Primers were designed to amplify a 180 bp PCR product. Figure 4-3 shows that tumour specific rearrangement was verified as somatic in the primary tumour and can be detected in whole genome amplified cfDNA. The next step is to quantify the amount of tumour mutation in this sample as a proof of principal that this approach could be used as a biomarker to monitor tumour burden in pancreatic cancer.

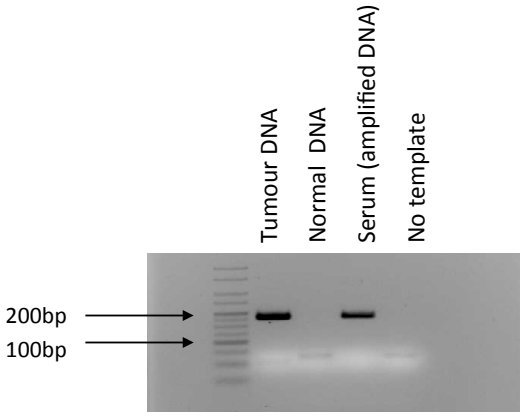


Figure 4-3 Gel photo of PCR validation of a deletion event in patient APGI 1959. Lane 1: Marker; Lane 2: Primary tumour DNA of patient APGI 1959 (positive control); Lane 3: Matched normal DNA (negative control); Lane 4: Whole genome amplified serum DNA; Lane 4: No template (negative control). (Data obtained from Nones et al., 2011)

Quantification of ctDNA using the tumour specific rearrangement

The Fluidigm BioMark quantification of the rearrangement (deletion) in the serum from patient APGI 1959 revealed that there were only 6 copies in the undiluted WGA cfDNA (24 ng/reaction) (Figure 4-4). No positive reaction was detected in the samples with further dilutions of the WGA cfDNA. The analysis of the undiluted WGA cfDNA was also tested on QX100™ Droplet Digital PCR System generating same results with 6 positive copies of tumour specific mutant detected (Supplementary Table 4-5).

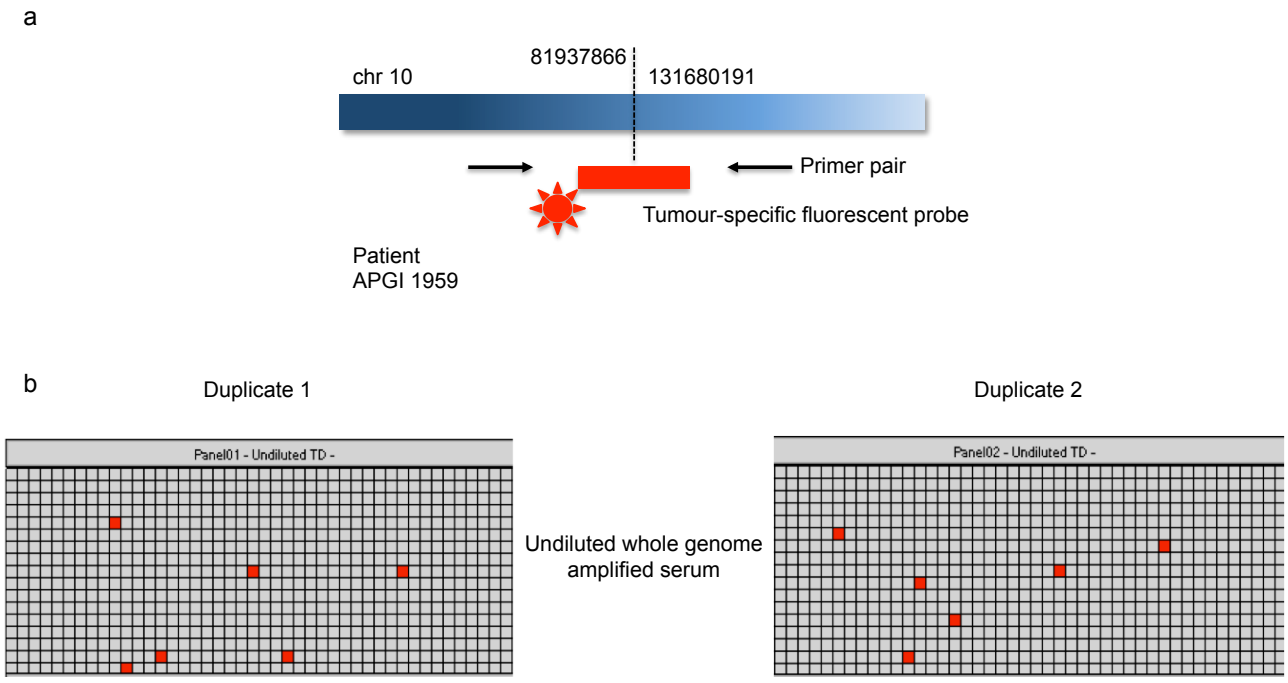


Figure 4-4 Analysis of ctDNA using Fluidigm BioMark System. (a) An illustration of TaqMan assay designed for dPCR. This is deletion event identified through whole genome sequencing of primary tumour from patient APGI 1959. With the knowledge of exact breakpoint, primers (black arrows) and fluorescent probes (red block) were designed to target the region spanning across the breakpoint position at chr10:81,937,866-131,680,191. (b) Representative heat map of digital array panel - each square represents a PCR reaction. When a tumour specific rearrangement is amplified, signals from the probe can be detected indicates in red.

4.3.2 Generic analysis – recurrent *KRAS* mutation

Detection limit of sequencing based approach to detect KRAS mutation

To determine the limit of detection of *KRAS* mutation, I compared sequencing results obtained from Ion Torrent PGM and Illumina MiSeq sequencers. The patient used for this experiment was APGI 1953 which contained a *KRAS* mutation (C>T) at position chr12:2539828. Serial dilution of tumour and matched normal DNA (20%, 4%, 0.8%. 0.16%, 0.032%, 0.0064%, 0.000128% of tumour DNA) was prepared in a total of 50 ng. A round of pre-amplification was performed (nested PCR) on each dilution. Sequencing was performed on each serial dilution of the PCR products to an average fold coverage of 350,000x (Ion Torrent) and 600,000x (MiSeq).

Ion Torrent PGM: To estimate the sequence error rate at this position, deep *KRAS* sequencing of germline DNA was performed. The estimated error rate is dependent on individual experiment and may vary between sequencing experiments and also at different positions in the genome. Here, the error rate was estimated to be 0.7% as the base change (C>T) was detected in 0.7% of the sequence reads in the germline DNA (Table 4-2). Using the error rate as a benchmark, I identified that sample containing 0.8% tumour DNA (expected allele frequency of 0.3%) is the limit of detection for this patient at read depth of 350,000x.

Illumina MiSeq: The detection limit of *KRAS* mutation was assessed in a similar manner that was described in PGM sequencing. Sequencing was performed on the MiSeq to a higher average read depth of 600,000x. The sequence error rate estimated by the *KRAS* mutation frequency in germline DNA was 0.3% (Table 4-2). These results suggest that MiSeq has a lower error rate and might be more applicable to cfDNA studies than Ion Torrent. However, in serial dilutions, the frequency of detected *KRAS* mutation had reached the error rate limit at sample containing 0.8% of tumour DNA (with expected allele frequency of 0.3%).

Table 4-2 Detection limit of *KRAS* detection using PGM and MiSeq sequencing

		Reads count for each base						
% of tumour DNA	**Amount of tumour DNA (corrected based on cellularity)**	**A**	**C**	**G**	**T**	**# count**	**Frequency mutation detection (%)***	**Expected % of mutation detection^**
PGM								
Germline DNA (0%)	0 ng	1337	342736	1544	2411	348028	0.7%	0.0%
Tumour DNA (100%)	6.9 ng	1225	184717	1678	158976	346596	46.3%	34.5%
1:5 (20%)	6.9 ng	1013	329113	602	24009	354737	6.8%	6.9%
1:25 (4%)	1.4 ng	1766	400134	2352	11627	415879	2.8%	1.4%
1:125 (0.8%)	0.276 ng	1226	325423	1672	3518	331839	1.1%	0.3%
1:625 (0.16%)	0.0552 ng	1060	361852	651	526	364089	0.1%	0.0552%
1:3,125 (0.032%)	0.011 ng	742	273383	521	813	275459	0.3%	0.0110%
1:15,625 (0.0064%)	0.0022 ng	1226	348138	1561	2070	352995	0.6%	0.0022%
1:78,125 (0.000128%)	0.0004 ng	1020	357058	638	644	359360	0.2%	0.0004%
MiSeq								
Germline DNA (0%)	0 ng	37	499679	50	1323	501089	0.3%	0.0%
Tumour DNA (100%)	6.9 ng	110	245629	180	194257	440176	44.2%	34.5%
1:5 (20%)	6.9 ng	63	544385	118	41362	585928	7.1%	6.9%
1:25 (4%)	1.4 ng	68	626132	70	9387	635657	1.5%	1.4%

1:125 (0.8%)	0.276 ng	44	515776	48	2436	518304	0.5%	0.3%
1:625 (0.16%)	0.0552 ng	74	645374	66	1874	647388	0.3%	0.0552%
1:3,125 (0.032%)	0.011 ng	85	721035	80	2263	723463	0.3%	0.0110%
1:15,625 (0.0064%)	0.0022 ng	75	695589	76	1831	697571	0.3%	0.0022%
1:78,125 (0.000128%)	0.0004 ng	73	708434	56	1609	710172	0.2%	0.0004%

Summary table is based on serial dilution of tumour DNA in germline DNA from patient APGI 1953 with cellularity of 69%. Mutation was detected on KRAS exon 2, chr12 position 25398284, C>T, aa: G12D

*Frequency mutation detection is the percentage of detected mutant molecules in a background of wild-type DNA (sum of reads counts of the base change (T) divided by the sum of total reads counts at the target position (C+T) x 100%)

^Expected percentage of mutation detection is calculated by taking into account of the cellularity and dilution of the analysed samples (cellularity of the analysed sample x tumour variant allele frequency 50% x dilution factor)

Previously, it has been shown that sequencing to detect tumour specific mutations in cfDNA is limited by tumour content in the sample, sequencing error, the coverage and the type of base substitution (Forshew et al., 2012). In an attempt to increase performance of *KRAS* by MiSeq sequencing, an additional patient (APGI 2353) which contained a dinucleotide base mutation of *KRAS* on exon 3 at codon 60 and 61 (GA>TT) was tested. Sequencing was performed to a higher average read depth of 1,000,000x. The sequence error rate was estimated at 0.066% using the *KRAS* mutation frequency in germline DNA (Table 4-3). As expected, it reduced the error rate and resulted in an improvement in our limit of detection. The frequency of detecting *KRAS* mutation reaches the error rate limit at sample containing 0.16% of tumour DNA (with expected allele frequency of 0.12%). Note that the primary tumour sample of this patient had much lower tumour content (cellularity of 30%). This is another factor that influences the detection of the analysis and needs to be considered with caution. The results cannot be directly compared to other runs as conditions of the runs were varied (for example, coverage, tumour content and the type of base interrogated). Unfortunately, the assay cannot be repeated due to the limitations in the amount of DNA available.

Based on the sequencing results, the error rates observed for benchtop sequencers at the *KRAS* position were 0.7% for PGM with an average read depth of 350,000x and 0.3% for MiSeq with read depth of 600,000x. These results suggest that the MiSeq sequencer would be a more suitable instrument for quantifying mutation load in circulating cfDNA as the amount of tumour DNA in total population of cfDNA is usually extremely low in the blood (Diehl et al., 2005).

Table 4-3 Detection limit for *KRAS* mutation using MiSeq sequencing on a patient with a dinucleotide mutation

		Reads count for double base change																	
DNA mixture (% of tumour DNA)	**Amount of tumour DNA (corrected based on cellularity)**	**AA**	**AC**	**AG**	**AT**	**CA**	**CC**	**CT**	**GA**	**GC**	**GG**	**GT**	**TA**	**TC**	**TG**	**TT**	**# count**	**Frequency mutation detection (%)***	**Expected % mutation detection^**
Germline DNA (0%)	0 ng	592	1	1	0	13	0	0	940229	193	1770	413	621	5	4	617	944474	0.066%	0%
Tumour DNA (100%)	3.0 ng	731	0	0	17	38	4	436	993971	245	1833	1154	1386	385	64	199360	1199643	16.706%	15.0%
1:5 (20%)	3.0 ng	708	1	0	2	44	3	102	1112327	241	2118	685	1055	76	17	48888	1166289	4.210%	3.0%
1:25 (4%)	0.6 ng	712	0	4	1	38	1	28	1133148	243	2272	557	807	17	5	9650	1147513	0.844%	0.6%
1:125 (0.8%)	0.12 ng	754	0	2	0	27	1	5	1291313	332	2476	614	799	10	4	3405	1299768	0.263%	0.12%
1:625 (0.16%)	0.024 ng	810	0	0	0	34	0	2	1210701	307	2454	571	798	5	4	1384	1217094	0.114%	0.024%
1:3,125 (0.032%)	0.0048 ng	776	0	0	1	48	1	3	1333446	305	2447	645	847	4	5	1051	1339614	0.079%	0.0048%
1:15,625 (0.0064%)	0.00096 ng	768	1	0	0	39	1	0	1172107	304	2239	630	780	5	7	947	1177846	0.081%	0.0010%
1:78,125 (0.000128%)	0.000192 ng	580	0	0	0	30	1	2	897109	180	1739	376	590	2	4	601	901237	0.067%	0.00019%

Summary table is based on serial dilution of tumour DNA in germline DNA from patient APGI 2353 with cellularity of 30%. Mutation was detected on KRAS exon 3, chr12 position 25380277- 25380278, GA>TT, aa: Q61K

*Frequency mutation detection is the percentage of detected mutant molecules in a background of wild-type DNA (sum of reads counts of the base change (T) divided by the sum of total reads counts at the target position (C+T) x 100%)

^Expected percentage of mutation detection is calculated by taking into account of the cellularity and dilution of the analysed samples (cellularity of the analysed sample x tumour variant allele frequency 50% x dilution factor)

Performance of PCR based approach to detect KRAS mutation

Digital PCR (dPCR) has been suggested to have higher sensitivity and precision than real-time quantitative PCR (Baker, 2012). Therefore, in an attempt to improve sensitivity of *KRAS* detection, two different dPCR instruments were evaluated - the Fluidigm BioMark system (chip-based) and Bio-Rad QX100 droplet digital PCR system (droplet-based).

Fluidigm BioMark System

Tumour and matched normal DNA from patient APGI 2106 was used to prepare the serial dilutions. Patient APGI 2106 contained a *KRAS* mutation (C>G) at position chr12: 25398285. In order to determine the optimal input amount of DNA for dPCR analysis in Fluidigm BioMark System, I used various amount of input DNA ranging from 5 ng to 50 ng. Duplicates were performed for 10 ng and 15 ng and showed a relatively consistency of detecting mutant DNA molecules between 0.8% and 20% of tumour DNA (10 ng: Pearson correlation coefficient = 0.99, $p<0.0001$; 15 ng: Pearson correlation coefficient = 0.98, $p=0.0002$). It was also found that the total amount of input DNA of 50 ng overloaded the system for Fluidigm dPCR analysis (data not shown), this is due to the number of partitions in a chip and thus, limits the volume of analysed sample.

In this case, the detection limit could not be determined as we do not have enough normal DNA to include as control. However, a positive relationship can be observed between the quantity of positive mutant molecules determined using Fluidigm BioMark system and the expected frequency of mutant molecules in each serial dilution across the different total amount of input DNA (5 ng: $R^2 = 0.99762$, 10 ng: $R^2 = 0.99811$ (avg.), 15 ng: $R^2 = 0.99181$ (avg.), 30 ng: $R^2 = 0.99333$) (Figure 4-5).

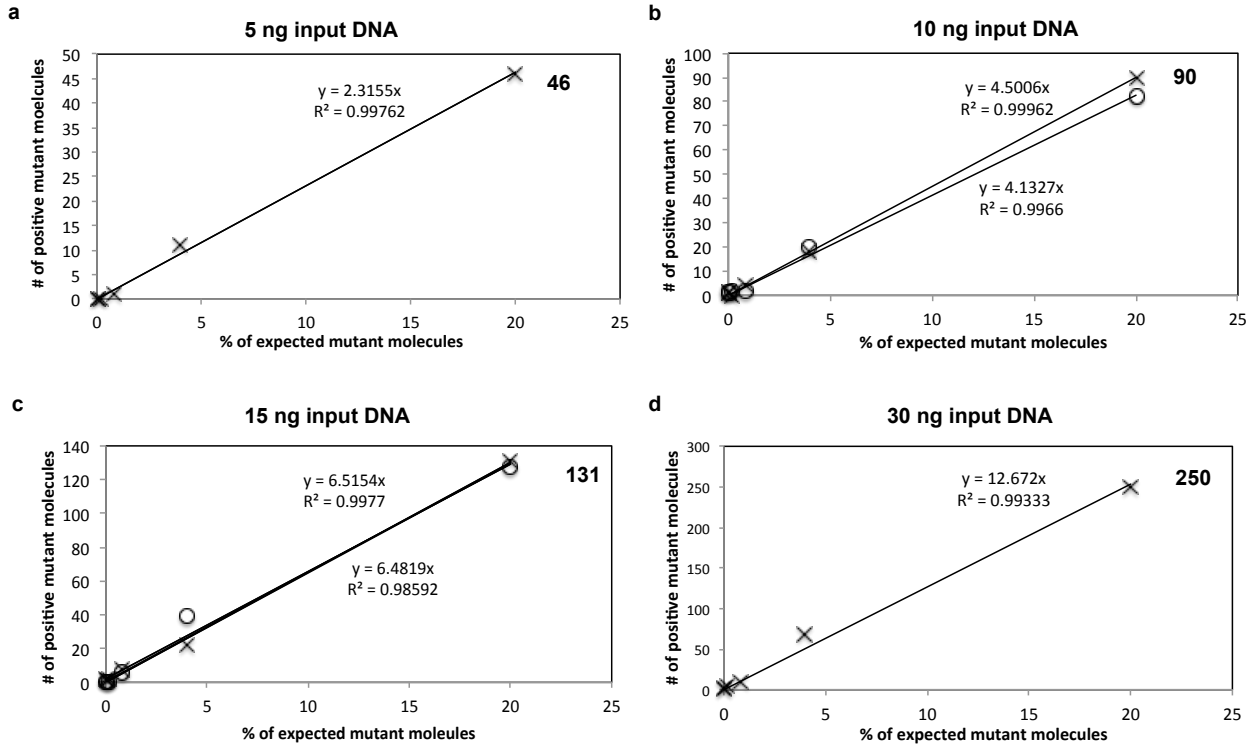


Figure 4-5 Performance of Fluidigm dPCR for the detection of *KRAS* mutation. Different total amount of input DNA mixture were analysed with dPCR – (a) 5ng total DNA, (b) 10 ng total DNA, (c) 15 ng total DNA, (d) 30 ng total DNA. The number of positive mutant molecules measured by dPCR assay is on the y-axis and the percentage of expected mutant molecules is on the x-axis. The number stated at the right corner of the plot is the maximum number of detected mutant molecules. The assay was designed to detect C>G base change at chr12: 25398285 (aa: G12R) in patient APGI 2106, which tumour DNA has 72.5% cellularity. The expected percentage of mutation detection is calculated by taking account of the cellularity and heterogeneity of the analysed sample (cellularity of the analysed sample x tumour variant allele frequency 50%). 10 ng and 15 ng input DNA were performed in duplicates.

QX100 droplet digital PCR system

An experiment was also performed using emulsion-based instrument known as droplet digital PCR (ddPCR) which allows partition of reactions as droplets. The instrument provides a higher number of reactions (~20,000) than a chip-based instrument (765 reactions). This higher number of reactions increases the dynamic range of emulsion PCR and allows accurate measurement of greater range of copies of alleles (Heyries et al., 2011; Pinheiro et al., 2012). Thus, this instrument also allows a higher input of DNA that might be beneficial to cfDNA studies.

A serial dilution experiment was performed in replicates of 30 ng and 60 ng total input DNA from patient APGI 1953. The number of mutant DNA molecules in both 30 ng and 60 ng input DNA was positively related to the estimated frequency of mutated allele in the serially diluted DNA mixture (Figure 4-6). The number of positive counts at 4% tumour DNA is 69 (Column – Run 1) when using total amount of 30 ng input DNA. By increasing the total amount of input DNA to 60 ng, the positive counts at 4% tumour DNA has achieved 136 (Column – Run 1). This may suggest that the rate of mutation detection increased when the amount of input DNA is higher (Figure 4-6, Supplementary Table 4-6)

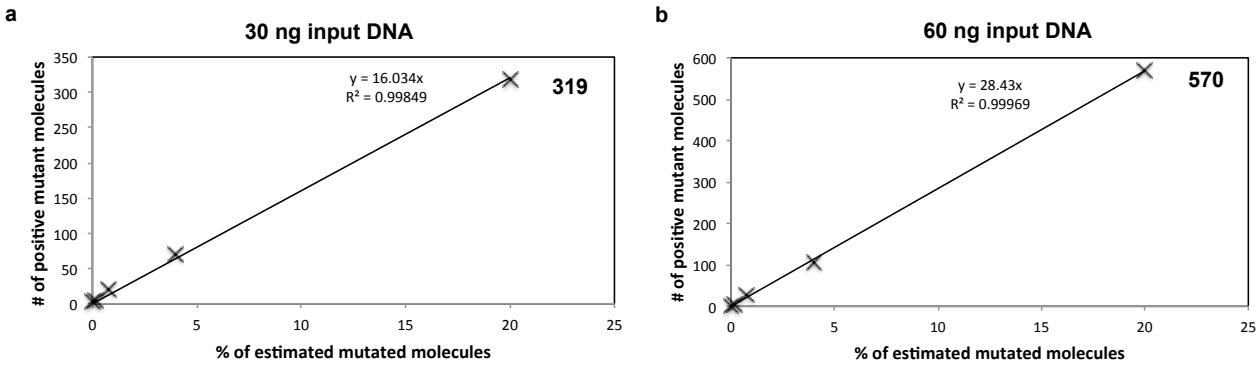


Figure 4-6 Performance of Bio-Rad ddPCR for the detection of *KRAS* mutation. Different total amount of input DNA mixture were analysed with dPCR – (a) 30ng total DNA, (b) 60 ng total DNA. The number of positive mutant molecules measured by dPCR assay is on the y-axis and the percentage of expected mutant molecules is on the x-axis. The number stated at the right corner of the plot is the maximum number of detected mutant molecules. Note that the graphs were fully based on a complete run that successfully detected normal and mutant molecules in all serial dilutions. And, the assay was designed for the detection of C>T base change (aa: G12D) in patient APGI 1953. The expected percentage of mutation detection is calculated by taking account of the cellularity and heterogeneity of the analysed sample (cellularity of the analysed sample x tumour variant allele frequency 50%).

Quantification of ctDNA using recurrent KRAS mutation

To determine if *KRAS* mutations are detectable in ctDNA, material from two patients was tested (APGI 1953 and APGI 2353).

Patient APGI 1953: This patient was tested for *KRAS* mutation of exon 2 of codon 12 (aa: G12D). The sequencing error rate obtained from the serial dilution of PGM sequencing experiment was 0.7% (Table 4-5). The assay was unable to detect the *KRAS* mutation in amplified plasma collected during post-operation and during chemotherapy. In this case, it suggests that the amount of ctDNA was less than 0.7%. Note that due to the availability of plasma, the quantification of plasma could not be performed on MiSeq or dPCR assay.

Table 4-4 Summary of *KRAS* detection in plasma of patient APGI 1953

			Reads count for each base						
Description	**DNA mixture (% of tumour DNA)**	**Amount of tumour DNA (corrected based on cellularity)**	**A**	**C**	**G**	**T**	**# count**	**Frequency mutation detection (%)***	**Expected % of mutation detection^**
*KRAS*_Nested	Normal DNA (0%)	0 ng	1337	342736	1544	2411	348028	0.7%	0.0%
*KRAS*_Nested	Tumour DNA (100%)	6.9 ng	1225	184717	1678	158976	346596	46.3%	34.5%
KRAS_Nested_cfDNA	Post-operation	-	1238	268359	1462	1824	272883	0.7%	NA
KRAS_Nested_cfDNA	During chemotherapy	-	840	307275	489	434	309038	0.1%	NA

Summary table is based on serial dilution of tumour DNA in germline DNA from patient APGI 1953 with cellularity of 69%. Mutation was detected on KRAS exon 2, chr12 position 25398284, C>T, aa: G12D

*Frequency mutation detection is the percentage of detected mutant molecules in a background of wild-type DNA (sum of reads counts of the base change (T) divided by the sum of total reads counts at the target position (C+T) x 100%)

^Expected percentage of mutation detection is calculated by taking into account of the cellularity and dilution of the analysed samples (cellularity of the analysed sample x tumour variant allele frequency 50% x dilution factor)

Patient APGI 2353: This patient contained a *KRAS* mutation in exon 3 of codon 60 and 61 (aa:Q61K). Three cfDNA samples were collected during the patients disease course (pre-operation, post operation 2nd day, and post operation 7th day). In all cases, the amount of cfDNA was limited, therefore the quantification assay was carried out using the two dPCR methods (i.e. Fluidigm and BioRad) as sequencing requires a high amount of DNA. Across the three cfDNA samples collected during the patients disease course (pre-operation, post operation 2nd day, and post operation 7th day), both dPCR assays showed no detection of mutant molecules in WGA plasma (data not shown).

Multiple rounds of PCR are often performed to improve the sensitivity of mutation detection in the serums (Forshew et al., 2012). To increase the sensitivity, an extra round of 18 PCR cycles was performed on the WGA plasma DNA targeting *KRAS* region prior to

the quantification using dPCR assays. The PCR amplified WGA plasma was then quantified in five replicates on two separate dPCR assays using both Fluidigm and Bio-Rad digital PCR system (data not shown). Even after an extra round of amplification, no mutant DNA molecules was detected in the population of circulating cfDNA suggesting that no mutant molecules was present in these samples or the level of mutant DNA were below the limits of detection of the instrument. In conclusion, no mutant DNA molecule was detected in plasma of patient APGI 2353 either in whole genome amplified samples or after multiple rounds of PCR.

Overall, the quantification results either using sequencing or dPCR assays re-emphasized the conclusion that there was little or no ctDNA present in the serum or plasma of patients APGI 1959, APGI 1953, and APGI 2353 and therefore hampered reliable detection.

4.3.3 Level of CA 19.9 and tumour burden

Here, I examined the timeline of these serum or plasma collected and the level of CA 19.9 across disease stages of the three pancreatic cancer patients. CA 19.9 is the most widely used and investigated marker for pancreatic cancer. Elevated levels of CA 19.9 are often associated with advanced pancreatic cancer or an indication of recurrent disease. Notably, patient APGI 1959 presented the highest level of CA19.9 with an average of ~567 U/mL ranging from 39 to 2520 U/mL suggesting that this patient might have the highest tumour burden as compared to the other two patients (Supplementary Table 4-7). The level of CA 19.9 of patient APGI 1953 was an average of ~82 U/mL ranging from 4 to 472 U/mL while average level of CA 19.9 of patient APGI 2352 was ~117 U/mL ranging from 16 to 267 U/mL. It was noted that the level of CA 19.9 of patient APGI 1959 was about 4 – 7 times higher than the other two patients. Hence, this suggests that the detection and quantification of tumour specific mutation in the serum from patient APGI 1959 might be due to the high tumour burden. Furthermore, the tumour specific mutation was detected in serum, which was collected after the spread of cancer to the liver (Figure 4-7). In contrast, plasma samples obtained from patients APGI 1953 and APGI 2353 were collected before the metastasis (Table 4-5, Figure 4-7). This observation suggests the detected tumour specific rearrangements (somatic deletion) in the circulating cfDNA of patient APGI 1959 could be derived from the liver metastasis sharing the same mutation with the primary pancreas tumour. Additionally, it might also suggest that the detection of mutant DNA molecules in the serum is more feasible in patients with metastatic disease.

Table 4-5 Summary of the analysed serum

Patient	Mutation (Position)	Cellularity of the primary tumour	Date of the serum collected	Status of the serum collection	Estimated level of CA 19.9 (U/mL)	# of mutant molecules detected by dPCR	Frequency of mutation detection by next generation sequencing
APGI 1959	Deletion (chr10:81,937,866-131,680,191)	83%	16/01/10	After the diagnosis of liver metastasis	~248-387	6	Did not perform
APGI 1953	*KRAS*, C>T (chr12: 25,398,284)	69%	29/12/10	Post-op	~16-65	Did not perform	0.7 % (PGM)
			04/01/11	During chemotherapy	~16-65	Did not perform	0.1% (PGM)
APGI 2353	*KRAS*, GA>TT (chr12:25,380,277- 25,380,278)	30%	03/06/11	Pre-op	~254-267	0	Did not perform
			06/06/11	Post-op, 2nd day	~254-267	0	
			10/06/11	Post-op, 7th day	~254-267	0	

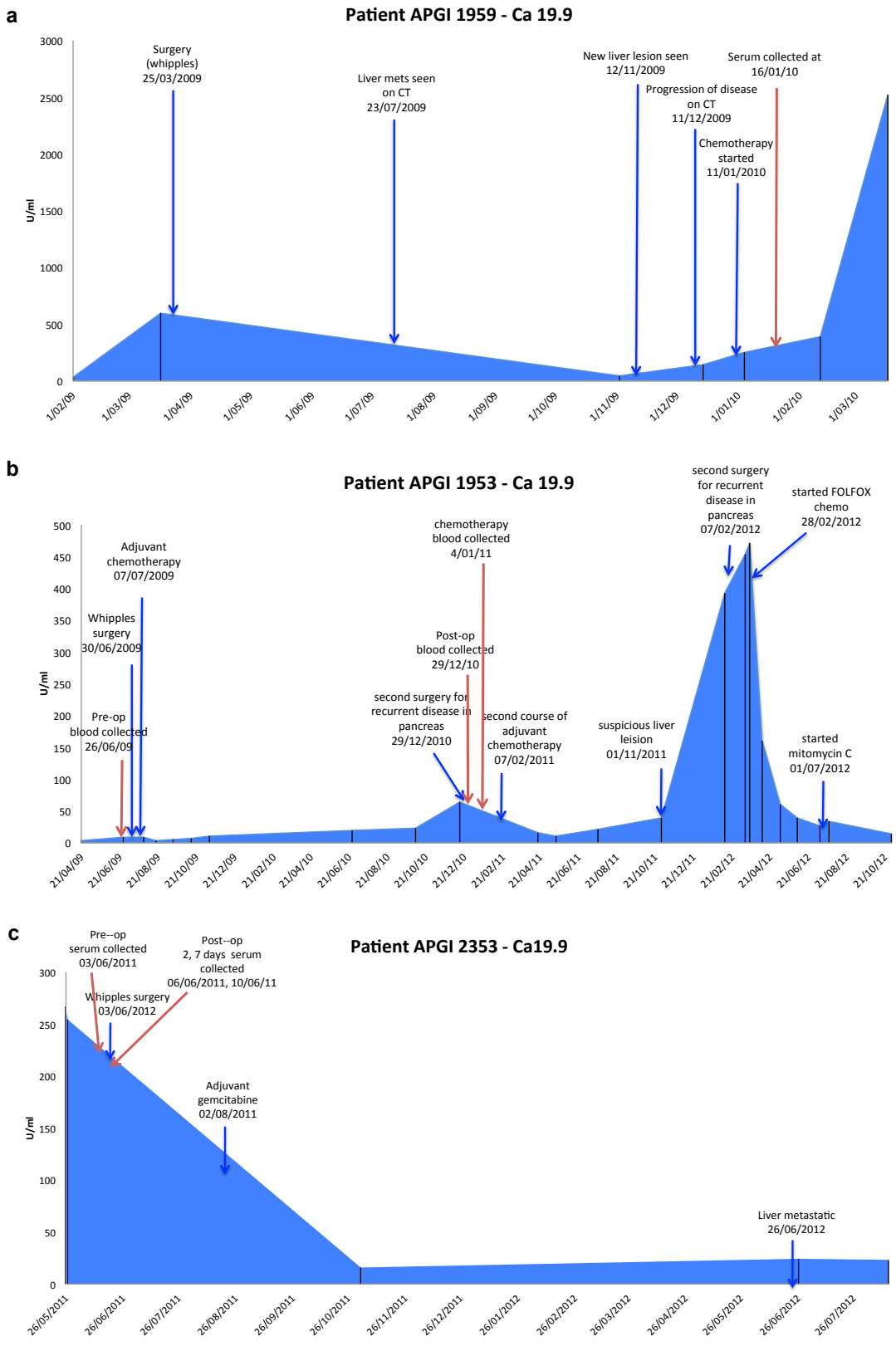


Figure 4-7 CA 19.9 levels across the 3 patients APGI 1959, APGI 1953, and APGI 2353. Arrow and text in red indicate the date of serum collected. Arrow in blue indicates the treatment and disease status of individual patient. Text in green indicates metastasis of the disease. The line in the blue area of the graph indicates the timepoints of CA 19.9 measurements. Supplementary Table 4-7 listed the value of CA 19.9 level.

4.4 Discussion

The evidence of tumour-derived mutations within circulating cfDNA was reported decades ago (Schwarzenbach et al., 2011; Sorenson et al., 1994; Vasioukhin et al., 1994). The quantification of cfDNA could effectively distinguish between patients with tumours and healthy individuals or monitoring patients after surgery and during chemotherapy. Studies carried out by Leary et al. have successfully detected tumour specific mutations in serum or plasma and have shown their potential as personalized biomarkers to monitor cancer disease progression in breast, bone and colorectal (Leary et al., 2010; Leary et al., 2012).

In this chapter, I investigated the possibility of using tumour specific mutations as clinical biomarkers for pancreatic cancer patients to improve the detection of recurrence and therapy response. I used two approaches: 1) a personalised approach using somatic rearrangement and 2) a generic approach using recurrent mutation of *KRAS* gene that is identified in most pancreatic cancers.

Personalised approach – somatic rearrangement

The quantification of ctDNA using tumour specific rearrangement was tested on serum of patient APGI 1959 using digital PCR (Fluidigm and Bio-Rad) and showed positive detection of 6 mutant copies. Even though the level of CA 19.9 was extremely high and the patient had metastatic disease, the amount of ctDNA in the circulation was very low. Unfortunately, due to the availability of serum, we could not repeat the assessment using next generation sequencing to validate the positive detection.

Generic approach – recurrent KRAS mutation

Two different methods (next generation sequencing and digital PCR) were assessed to determine the detection limit of *KRAS* mutation using serial dilutions of tumours. The results showed that both benchtop sequencers (PGM and MiSeq) and digital PCR assays (Fluidigm and Bio-Rad) have the capacity to detect tumour specific mutation in serial dilutions of tumour. PGM could achieve a detection limit for mutant allele fraction at 0.7% with read depth of 350,000x while the detection limit of MiSeq was at 0.3% with a higher deep coverage of 600,000x. The result from MiSeq is within the range of frequency mutation detection limit of what have been previously reported by numerous studies (0.02 - 2%) (Forshew et al., 2012; Narayan et al., 2012; Newman et al., 2014). The detection limits are varied across the different studies as they used different sequencing depth, algorithm as well as different cancer types was being analysed. However, based on the

results presented here, MiSeq would be more suitable for the quantification for *KRAS* mutation in ctDNA. However in this project, the MiSeq machine was acquired later, thus the quantification of ctDNA in patient APGI 1953 which initially performed using PGM, could not be repeated with the MiSeq as there was no extra DNA or plasma available.

With respect to dPCR, the detection limit could not be determined as we do not have enough normal DNA to include as control. However, the droplet dPCR (Bio-Rad) assays have shown that mutation detection increased when the amount of DNA input is higher. In contrast, the chip-based dPCR (Fluidigm) was limited by the input DNA attribute to the number of partitions and thus, lesser volume of the samples could be analysed. Regardless, both chip (Fluidigm) and droplet based (Bio-Rad) dPCR showed linear relationship with the expected frequency of mutation detection indicating the ability of these assays to quantify mutations in a single experiment.

During the quantification of ctDNA in patients APGI 1953 and APGI 2353, no *KRAS* mutation was detected in the cfDNA from plasma. Together with the result obtained from APGI 1959, it suggests that there was little or no tumour specific mutation in the serum or plasma across these three patients. The reasons why ctDNA was detected in patient APGI 1959 and not patients APGI 1953 and patient 2353 could be due to the high CA19.9 level and the late stage disease when serum or plasma were collected. It was noted that the plasma collected from patients APGI 1953 and APGI 2353 were at the early stage of the disease (before liver metastasis) and these patients had a lower CA19.9 level with an average of ~82 and ~117 U/mL. In contrast, the serum collected from patient APGI 1959 was at the later stage of the disease (presenting liver metastasis) and the patient had an extremely high CA19.9 level with an average of ~567 U/mL. The high level of CA19.9 suggests higher tumour burden in patient APGI 1959, therefore increased the likelihood of mutation detection rate in the circulation. However, it could also be possible that the detected ctDNA in patient APGI 1959 may be derived from the liver metastasis, which shared the same mutation with the primary pancreas tumour.

Potential factors that may affect detection of tumour DNA in circulation and might explain why no ctDNA was detected in the two patients. 1) Nature of pancreatic tumour. Pancreatic tumours generally contain a large amount of stromal and fibrotic tissue, which results in low blood perfusion and a hypoxic environment, therefore reducing the chance of ctDNA to be released into the blood (Yu and Tannock, 2012). 2) The tissue structure of

pancreas. Fragmented tumour DNA leaving the pancreas may get processed and broken down by the liver before entering the bloodstream (Nussey and Whitehead, 2001), therefore reducing the ability to detect tumour specific mutation in the circulation. 3) Disease stage. A recent study carried out by Bettegowda et al. (Bettegowda et al., 2014) analysed the detection of ctDNA in early and late stage of human cancers including pancreatic cancer in a large cohort of 640 patients. They reported that the detection rate of ctDNA in the blood samples was correlated with the disease stage of cancer and the concentration of ctDNA increases as the disease progressed. From the 34 plasma samples from pancreatic cancer, the authors detected ctDNA in approximately 90% of the cases with metastatic disease, and only approximately 50% of the cases with localised pancreatic cancer. Their findings suggest that at this current stage, the detection of ctDNA is more feasible in patients with metastatic disease, which agree with the results presented in this chapter. In the context of pancreatic cancer, the quantification might be more suitable to test recurrent disease, which metastasized away from the pancreas. 4) The volume of serum or plasma used to extract cfDNA may be also crucial. Bettegowda et al. (Bettegowda et al., 2014) used 5 mL of plasma, which is double the amount of plasma or serum used in this study.

Furthermore, this project also contained several logistic limitations. These include genomic contamination of several serum DNA samples, the slow progress of recruitment of new patients with ctDNA and the need to draw more blood from patients.

4.5 Supplementary Material

Supplementary Table 4-1 Summary of serum or plasma DNA received

Patient	Garvan ID	Type	Volume of initial serum/plasma (uL)	Volume received (uL)	Quantity (ng /uL)	Quality
APGI-1830	8048277	Post op	955	20	0	Fail
APGI-1830	8048278	Post op	955	<10	0	Fail
APGI-1830	8048276	Post-op	955	20	0	Fail
APGI-1953	8047164	During chemotherapy	900	20	0.89	Pass
APGI-1953	8047172	Post-op	900	20	0.65	Pass
APGI-1953	8047973	Pre-op	840	20	0	Fail
APGI-1953	8047168	Pre-op	430	20	0	Fail
APGI-1953	8048813	Pre-op	900	25	0	Fail
APGI-1953	8047974	Pre-op	885	25	0	Fail
APGI-1953	8047165	During chemotherapy	900	8	0.32	Pass
APGI-1953	8047173	Post-op	900	12	0.14	Pass
APGI-2302	8047990	Post-op	840	20	2.3	Fail
APGI-2302	8047993	Post-op	840	20	1.01	Fail
APGI-2302	8047976	Pre-op	840	20	0	Fail
APGI-2302	8047976	Pre-op	1755	20	0	Fail
APGI-2353	8048285	Post-op (2 days)	1755	20	5	Pass
APGI-2353	8048286	Post-op (2 days)	1755	20	1.24	Pass
APGI-2353	8048287	Post-op (2 days)	1755	<10	5.3	Pass
APGI-2353	8048298	Post-op (60 days)	1755	20	0.3	Pass
APGI-2353	8048300	Post-op (60 days)	1755	<10	0.3	Fail
APGI-2353	8048299	Post-op (60 days)	1755	20	0	Fail
APGI-2353	8048291	Post-op (7 days)	1755	20	1.62	Pass
APGI-2353	8048292	Post-op (7 days)	1755	20	2.44	Pass
APGI-2353	8048293	Post-op (7 days)	1755	<10	2.33	Pass
APGI-2353	8048279	Pre-op	1755	20	1.50	Pass
APGI-2353	8048280	Pre-op	1755	20	0.50	Pass
APGI-2353	8048281	Pre-op	1755	<10	1.58	Pass
APGI-1959	8016633	During chemotherapy	500	NA	NA	Pass

Pre-op: It refers to blood collected before operation.
Post op: It refers to blood collected after operation (indicates the number of days after operation).

Pass: It indicates that the size of the serum or plasma DNA lies within expected range between 130 and 200 bp. Fail: It indicates that the serum or plasma DNA is contaminated with genomic DNA that could due to cell lyse

Supplementary Table 4-2 Summary of the amount of serum or plasma DNA used throughout the attempt of quantification

Patient id	Description of the serum/plasma	Quantity of serum/plasma DNA received (volume)	Quantity of serum/plasma DNA used in whole genome amplification	Quantity of amplified serum/plasma DNA (volume)	Quantity of amplified serum/plasma DNA used per nested PCR reaction	Quantity of amplified material used per digital PCR run (Fluidigm & Bio-Rad)	Quantity of amplified material used per sequencing run (PGM & MiSeq)
APGI 1959	During chemotherapy	0.1 ng (NA)	NA	24 ng (1 uL)	NA	24 ng	NA
APGI 1953	Post-operation	13 ng (20 uL)	~ 9 ng	46.2 ng (14 uL)	2 ng	NA	10 ng
APGI 1953	During chemotherapy	17.8 ng (20 uL)	~ 9 ng	46.2 ng (14 uL)	2 ng	NA	10 ng
APGI 2353	Pre-operation	30 ng (20uL)	~ 20 ng	360 ng (15 uL)	2 ng	36 ng	NA
APGI 2353	Post-operation (after 2 day)	24.8 ng (20 uL)	~ 20 ng	381 ng (15 uL)	2 ng	36 ng	NA
APGI 2353	Post-operation (after 7 day)	32.4 ng (20uL)	~ 20 ng	435 ng (15 uL)	2 ng	36 ng	NA

Supplementary Table 4-3 Details of primers and barcodes used for *KRAS* mutation detection in next generation sequencing

Description	(Patient ID) Primer sequence	Barcode sequence for PGM	Barcode sequence for MiSeq
Nested PCR 1st round	(APGI 1953) F: 5'- ATTCGTCCACAAAATGATTC R: 5'- GTTCTAATATAGTCACATTTTCATT	>IonXpress_001 CTAAGGTAAC >IonXpress_002 TAAGGAGAAC >IonXpress_003 AAGAGGATTC >IonXpress_004 TACCAAGATC >IonXpress_005 CAGAAGGAAC >IonXpress_006 CTGCAAGTTC >IonXpress_007 TTCGTGATTC >IonXpress_008 TTCCGATAAC >IonXpress_009 TGAGCGGAAC >IonXpress_010 CTGACCGAAC >IonXpress_011 TCCTCGAATC >IonXpress_012 TAGGTGGTTC >IonXpress_013 TCTAACGGAC >IonXpress_014 TTGGAGTGTC >IonXpress_015 TCTAGAGGTC >IonXpress_016 TCTGGATGAC	>MID_1 ACGAGTGCGT >MID_2 ACGCTCGACA >MID_3 AGACGCACTC >MID_4 AGCACTGTAG >MID_5 ATCAGACACG >MID_6 ATATCGCGAG >MID_7 CGTGTCTCTA >MID_8 CTCGCGTGTC >MID_10 TCTCTATGCG >MID_11 TGATACGTCT >MID_13 CATAGTAGTG >MID_14 CGAGAGATAC >MID_15 ATACGACGTA >MID_16 TCACGTACTA >MID_17 CGTCTAGTAC >MID_18 TCTACGTAGC >MID_19 TGTACTACTC >MID_20 ACGACTACAG >MID_21 CGTAGACTAG >MID_22 TACGAGTATG >MID_23 TACTCTCGTG
2nd round	(APGI 1953) F: 5'- TATCGTCAAGGCACTCTTGC R: 5'- GCCTGCTGAAAATGACTGAA		
	(APGI 2353) F: 5'- TACTGGTCCCTCATTGCACTGT R: 5'- TGATGGAGAAACCTGTCTCTTGGA		

Supplementary Table 4-4 Details of primers and probes used for tumour specific mutations detection in dPCR

Patient ID	APGI 1959	APGI 1953	APGI 2353
Event type	Deletion	*KRAS* codon 12	*KRAS* codon 60, 61
Event size (bp)	50,000	NA	NA
Base change	NA	Codon 12; C>T	Codon 60, 61; GA>TT
Position	chr10:81,937,866-131,680,191	chr12: 25,398,284	chr12:25,380,277- 25,380,278
aa change	NA	G12D	Q61K
Primes sequence (5'-)	F: 5' - TCCGCTTGGTACAGGGAGAAGCA R: 5'-GGGCTCCAGTTAAAACGGTTGATCCC	F: 5'-GCCTGCTGAAAATGACTGAATATAAACT R: 5'-GCTGTATCGTCAAGGCACTCTT	F: 5'-CATGTACTGGTCCCTCATTGCA R: 5' - GATGGAGAAACCTGTCTCTTGGAT
Probes sequence	FAM 5'-TGGCCGGGAGAGTCCCAACAAAAAA	FAM 5'-TTGGAGCTGATGGCGTA VIC 5'-TTGGAGCTGGTGGCGTA	FAM 5'-TGTACTCCTCTTTTCCTGCTG VIC 5'-TGTACTCCTCTTTTCCTGCTG
Amplicon size	180 bp	78 bp	78 bp

Supplementary Table 4-5 Analysis of ctDNA using QX100™ Droplet Digital PCR System

Sample	Dilution	Data Quality	Concentration	Copies/rxn	Poisson Conf Min	Poisson Conf Max	Positives	Negatives
Primary tumour APGI 1959	1:100	0	0	0	0	0.278	0	11850
Serum APGI 1959	1:1	99.5	0.459	9.18	0.18	0.938	6	14368
	1:2	99.7	0.255	5.1	0.0606	0.677	3	12913
	1:5	0	0	0	0	0.503	0	6549
	1:10	0	0	0	0	0.262	0	12551

Supplementary Table 4-6 Summary of dPCR assay using QX100™ Droplet Digital PCR system

			VIC signal (Wild type)		FAM signal (Mutant)				
Total amount DNA	**DNA mixture**	**Tumour DNA per reaction**	**# Estimated targets^**		**# Estimated targets^**			**Mean of mutation detection (%)[§]**	**Expected % of mutation detection[¶]**
			Run 1	Run 2	Run 1	Run 2	Mean		
30 ng	1:5 (20%)	6.9 ng	4361	0*	319	0*	-	6.28%	6.9%
	1:25 (4%)	1.4 ng	4396	4215	71	69	70 ±1.4	1.60%	1.4%
	1:125 (0.8%)	0.276 ng	4871	3476	20	9	14.5 ±7.8	0.35%	0.3%
	1:625 (0.16%)	0.0552 ng	4859	4502	4	4	4 ± 0	0.082%	0.0552%
	1:3,125 (0.032%)	0.011 ng	4357	3723	3	1	2 ±1.4	0.069%	0.0110%
60 ng	1:5 (20%)	6.9 ng	472*	6890	34*	570	-	7.64%	6.9%
	1:25 (4%)	1.4 ng	8592	6577	136	106	121 ±21.2	1.57%	1.4%
	1:125 (0.8%)	0.276 ng	8629	7785	44	26	35 ±12.7	0.42%	0.3%
	1:625 (0.16%)	0.0552 ng	8169	7656	23	5	14 ±12.7	0.177%	0.0552%
	1:3,125 (0.032%)	0.011 ng	7328	7000	11	1	6 ±7.1	0.084%	0.0110%

Analysis table is based on DNA mixture of patient APGI 1953 with cellularity of 69%. Mutation was detected on KRAS exon 2, chr12 position 25398284, C>T using QX100™ Droplet Digital PCR system.

^Estimated targets estimated number of molecules present in a reaction determined by QX100™ Droplet Digital PCR software.

[§]Mean of mutation detection is the percentage of mutant molecules in a background of wild type DNA (sum of estimated targets for mutant molecules divided by sum of estimated targets for wild type and mutant molecules x 100%)

The expected percentage of mutation detection was calculated by taking into account with the cellularity, heterogeneity and dilution factor of the analysed samples (tumour variant allele frequency 50% x cellularity of the analysed sample x dilution factor)

*Failed run or errors generated during set up and preparation of the assay.

Supplementary Table 4-7 Levels of CA 19.9 of the 3 analysed patients across their clinical journey^

APGI 1959 – Date of measurement	Level (U/mL)
1/02/09	39
17/03/09	594
2/11/09	41
14/12/09	143
4/01/10	248
11/02/10	387
17/03/10	**2520**

APGI 2353 – Date of measurement	Level (U/mL)
26/05/11	**267**
27/05/11	254
2/11/11	16
26/06/12	24
14/08/12	23

APGI 1953 – Date of measurement	Level (U/mL)
21/04/09	4
26/06/09	9
30/07/09	8
18/08/09	4
14/09/09	5
13/10/09	7
10/11/09	10
26/06/10	20
5/10/10	23
15/12/10	65
18/04/11	16
16/05/11	11
23/07/11	21
1/11/11	39
9/02/12	393
12/03/12	453
20/03/12	**472**
9/04/12	161
7/05/12	61
4/06/12	40
9/07/12	26
24/07/12	34
1/11/12	14

^ Note that the level of CA 19.9 corresponds to Figure 4-5.

5 Chapter 5

Final conclusion and discussion

5.1 Summary

Throughout this thesis, I presented a comprehensive study of somatic rearrangements in primary pancreatic tumours and focussed on three different aspects: the exploration of genomic technologies and computational methods to detect and verify somatic rearrangements, analysis of somatic rearrangements and breakpoints, and the utility of tumour specific rearrangements and mutations in personalized genomic medicine.

In Chapter 2, I established a high throughput workflow to rapidly verify somatic rearrangements and identified exact breakpoints to base pair resolution in primary tumours using genomic and computational tools. **Highlights**: The advantages of this high throughput workflow include (i) the utilization of benchtop next generation sequencing (Ion Torrent PGM and Illumina MiSeq) to replace conventional Sanger sequencing, which mitigated the concern of PCR product size allowing primers to be designed for more events; (ii) the integration of bioinformatics tools and next generation sequencing based methods, which greatly increased the speed and volume of the verification process; and (iii) the accuracy of the next generation sequencing methods was comparable to the conventional Sanger sequencing method. Furthermore, this workflow is applicable to both long mate pair and pair end sequencing data as the qSV tool has the ability to identity structural rearrangements from SOLiD and Illumina sequencing platforms. This work has been published in BioTechniques Report as a method article in early 2014 (Quek et al., 2014).

In Chapter 3, I described the spectrum of somatic rearrangements and breakpoints detected in 120 primary pancreatic tumours. The analyses conducted include the characterization of breakpoint junctions, the examination of breakpoint patterns at the junctions using the information obtained from split contig alignment as well as searching for DNA enrichment patterns from the reference genome surrounding the breakpoint. **Highlights:** A total of 10,074 high confidence somatic rearrangements were identified from 120 pancreatic primary tumours. Pancreatic cancer is heterogeneous disease as each

pancreatic cancer genome displays distinct numbers and patterns of genomic breakpoints. The exact breakpoints were resolved in 95% of the somatic rearrangements. The majority of the breaks exhibited 74.8% microhomology and approximately 67.0% of these microhomology rearrangements contained short overlapping bases of 1 to 5 bp, which suggest that NHEJ might be the predominant DNA repair mechanism in pancreatic cancer. Pancreatic tumours associated with *BRCA* mutations and/or a high contribution of the BRCA mutational signature showed a higher frequency of somatic breakpoints with microhomology length of 1 to 5 bp and lower frequency of blunt end (0 bp) when compared to *BRCA* wild type or low BRCA mutational signature tumours. A similar pattern was also observed in a cohort of ovarian tumours. Together, these datasets suggest that different NHEJ pathways were utilised depending on the *BRCA* mutation status. Furthermore, the *BRCA* deficiency tumours that were associated with HR defective pathway seem to favour the restoration of the breaks by 'error-prone' NHEJ pathway. Thus, the identification of breakpoints with microhomology may indicate the process of defective HR pathway in a subgroup of pancreatic tumours and could response to PARP1 inhibitors.

In addition, the analysis of the DNA sequences surrounding the breakpoints revealed strong signals of A+Ts rich regions suggesting that the formation of somatic rearrangements in pancreatic cancer could be in part mediated by either retrotransposition activity or chromosomal fragile sites, however these findings need to be further investigated. Taken together, the results presented in Chapter 3 provide insights into the complexity of breakpoints patterns in pancreatic cancer genomes. The different composition of event types and differences between the breakpoint characteristics across the pancreatic cancer genomes may suggest that the somatic rearrangements in pancreatic tumorigenesis could be formed by multiple DNA repair mechanisms.

The categorization of thousands of chromosomal somatic rearrangements across 120 pancreatic primary tumours revealed that each cancer genome harbours a diversity of somatic structural rearrangements and mutations. Therefore, it was hypothesized that these tumour specific alterations could be used as potential cancer biomarkers for clinical utility.

In Chapter 4, I outlined a framework that combined tumour specific somatic rearrangements (personalized) and recurrent *KRAS* mutation (generic) to quantify ctDNA in serums or plasma collected from pancreatic cancer patients following a course of

therapy. The aim was to identify an alternative approach utilising verified somatic rearrangements and mutations as low invasive biomarkers in the serum or plasma of pancreatic cancer patients to trace the course of disease or during cancer treatment. This work could potentially show the use of next generation sequencing technology as a diagnostic tool in the clinic. **Highlights:** PCR based and sequencing based methods have demonstrated the capacity of detecting and quantifying ctDNA. In this subset of pancreatic cancer patients, the detection limits for sequencing and digital PCR methods was <1%. Unfortunately, there was little or no ctDNA detected in the serum or plasma samples from the cases studied. Perhaps due to the nature of pancreas cancer, the tissue structure of pancreas, disease progression, and volume of plasma and serum collected for DNA isolation have affected the quantification of tumour specific mutation in the circulation.

Overall, the results from Chapter 4 showed the following limitations and challenges: (1) the limited amount of cfDNA obtained from <2 mL of blood has imposed a challenge in quantifying ctDNA; (2) the area of circulating DNA is still at the early stage, the current techniques and technologies used for isolation and detection of ctDNA may have limited capacity and sensitivity to detect mutations that represent 0.01% of the total population of cell free circulating DNA; (3) most of the pancreatic cancer patients are diagnosed at late stage and have a median survival of 6 months, this increases challenge in recruiting more samples to follow the disease course in pancreatic cancer patients.

5.2 Future studies

The current analyses have provided a foundation, however, we could still improve the methods and analyses to further explore the data and better understand the process of somatic rearrangements formation in pancreatic carcinogenesis and the clinical utility of such rearrangements.

Chapter 2 – A High throughput workflow to rapidly detect and identified somatic rearrangements breakpoints

The high throughput workflow still required manual steps to confirm potential structural rearrangements. Therefore, the workflow can be improved by the following: (1) Elimination of the gel electrophoresis step. The PCR amplification of tumour and matched normal DNA from individual patient can be barcoded, pooled and subsequently sequenced on benchtop sequencing platforms. After the assembly, the somatic events can be determined by selecting the breakpoints identified in tumour and not in normal. Hence, this procedure can

further reduce the amount of time and labour for the analysis of somatic rearrangements. (2) Complete removal of wet lab verification such as PCR and gel electrophoresis. The evaluation of the sensitivity and accuracy of identifying somatic rearrangements in cancer can be validated by using multiple callers and different sequencing platforms (such as HiSeq X and NextSeq 500).

Chapter 3 – Characterisation and patterns of somatic breakpoints

This chapter has reported a comprehensive catalogue of somatic rearrangements and breakpoint characteristics of pancreatic cancer. The breakpoint characteristics have provided an insight into the potential operative mechanisms of generating somatic rearrangements. To further explore the formation mechanisms of somatic rearrangements, future studies could consider other genomic factors that might impact on the numbers and distribution of breakpoints in the genome and these were previously suggested by Drier et al. (Drier et al., 2013). Genomic factors include replication time, transcription rate and GC content. In their study, they have shown that these factors can differ by tumour types and the pattern of breakpoints might explain cellular processes that generate or repair the somatic rearrangements. The identification of breakpoints with a higher frequency of microhomology in *BRCA* mutated and high BRCA mutational signature samples suggests that this group of pancreatic tumours might respond to PARP inhibitors and/or DNA damaging agents. However to confirm this, further studies would need to be performed using patient material from those that did and did not respond to PARP inhibitors and/or DNA damaging agents.

The analysis of the DNA sequence surrounding the breakpoints indicates the signal of A+Ts enrichment in the regions around the breakpoints. Further analysis to pinpoint the potential mechanism is required so that we can clarify that A+Ts rich sequences are associated with retrotransposons and/or fragile sites. Analysis could include the study of the relative distance of A+Ts motifs from the exact breakpoint location or mapping the DNA sequences surrounding the breakpoints to repeats database and/or known fragile sites regions to further elucidate the mechanism. Furthermore, the DNA patterns at the breakpoints allow additional analysis such as methylation sites, histone marks and tandem repeats. All these analyses together may give us more clues on the operative mechanisms in pancreatic cancer.

Despite the slow recruitment of patients, here is some of the potential future work to address the limitations that I have encountered when detecting ctDNA in pancreatic cancer:

(1) Increase the starting volume of the blood to maximise the sensitivity of detecting ctDNA (as currently we were only collecting approximately <2 mL of blood) and the volume of blood withdraw should be consistent throughout the course of study.

(2) Improve isolation methods to isolate ctDNA. The use of circulating cfDNA is still at its early stage, as the field matures, apart from QIAamp DNA Blood Kit (Qiagen, Inc.) which is widely used for isolating circulating cfDNA from serum or plasma, it is possible that more commercial kits or improvements in existing techniques become available and bring more options to better isolate ctDNA.

(3) To improve rare allele enrichment, we could evaluate the COLD-PCR (co-amplification at lower denaturation temperature PCR) method. As suggested, COLD-PCR prior to sequencing or genotyping assays can improve the ability of detecting mutation by 50-fold (Milbury et al., 2012)

(4) The use of exosomes as biomarkers. Previous study has demonstrated that more than 10 kb fragments of genomic DNA spanning across *KRAS* and *TP53* mutations were identified in the serum exosomes of patients with pancreatic cancer (Kahlert et al., 2014). Exosomes are very stable in various conditions, and the genetic materials are protected against degradation and denaturation in extracellular environment by the vehicle (Taylor and Gercel-Taylor, 2008). Thus, exosomes might be a promising biomarker for pancreatic cancer.

5.3 Closing remarks

This is an exciting phase of cancer genomics as sequencing embarks on a new journey of how we think about cancer and cancer treatment. In this thesis, I presented a high throughput method to verify rearrangements, conducted analysis to explore factors that may give us clues about the differences in numbers and distribution of somatic rearrangements in pancreatic cancer and finally explored the clinical utility of these events in patients with pancreatic cancer.

Reference

(2010). International network of cancer genome projects. Nature *464*, 993-998.

(2011). Integrated genomic analyses of ovarian carcinoma. Nature *474*, 609-615.

Alexandrov, L.B., Nik-Zainal, S., Wedge, D.C., Aparicio, S.A., Behjati, S., Biankin, A.V., Bignell, G.R., Bolli, N., Borg, A., Borresen-Dale, A.L.*, et al.* (2013). Signatures of mutational processes in human cancer. Nature *500*, 415-421.

Alkan, C., Coe, B.P., and Eichler, E.E. (2011). Genome structural variation discovery and genotyping. Nature Reviews Genetics *12*, 363-376.

Argueso, J.L., Westmoreland, J., Mieczkowski, P.A., Gawel, M., Petes, T.D., and Resnick, M.A. (2008). Double-strand breaks associated with repetitive DNA can reshape the genome. Proceedings of the National Academy of Sciences of the United States of America *105*, 11845-11850.

Arlt, M.F., Durkin, S.G., Ragland, R.L., and Glover, T.W. (2006). Common fragile sites as targets for chromosome rearrangements. DNA Repair *5*, 1126-1135.

Baca, S.C., Prandi, D., Lawrence, M.S., Mosquera, J.M., Romanel, A., Drier, Y., Park, K., Kitabayashi, N., MacDonald, T.Y., Ghandi, M.*, et al.* (2013). Punctuated evolution of prostate cancer genomes. Cell *153*, 666-677.

Bailey, T.L., Boden, M., Buske, F.A., Frith, M., Grant, C.E., Clementi, L., Ren, J., Li, W.W., and Noble, W.S. (2009). MEME SUITE: tools for motif discovery and searching. Nucleic Acids Research *37*, W202-208.

Baker, M. (2012). Digital PCR hits its stride. Nature Methods *9*, 541-544.

Barlow, J.H., Faryabi, R.B., Callen, E., Wong, N., Malhowski, A., Chen, H.T., Gutierrez-Cruz, G., Sun, H.W., McKinnon, P., Wright, G.*, et al.* (2013). Identification of early replicating fragile sites that contribute to genome instability. Cell *152*, 620-632.

Bartek, J., and Lukas, J. The DNA damage response in tumorigenesis and cancer treatment. Nature Reviews Cancer.

Beaver, J.A., Jelovac, D., Balukrishna, S., Cochran, R.L., Croessmann, S., Zabransky, D.J., Wong, H.Y., Valda Toro, P., Cidado, J., Blair, B.G.*, et al.* (2014). Detection of cancer DNA in plasma of patients with early-stage breast cancer. Clinical Cancer Research *20*, 2643-2650.

Berger, M.F., Lawrence, M.S., Demichelis, F., Drier, Y., Cibulskis, K., Sivachenko, A.Y., Sboner, A., Esgueva, R., Pflueger, D., Sougnez, C.*, et al.* (2011). The genomic complexity of primary human prostate cancer. Nature *470*, 214-220.

Bettegowda, C., Sausen, M., Leary, R.J., Kinde, I., Wang, Y., Agrawal, N., Bartlett, B.R., Wang, H., Luber, B., Alani, R.M.*, et al.* (2014). Detection of circulating tumor DNA in early- and late-stage human malignancies. Science Translational Medicine *6*, 224ra224.

Biankin, A.V., Waddell, N., Kassahn, K.S., Gingras, M.C., Muthuswamy, L.B., Johns, A.L., Miller, D.K., Wilson, P.J., Patch, A.M., Wu, J.*, et al.* (2012). Pancreatic cancer genomes reveal aberrations in axon guidance pathway genes. Nature *491*, 399-405.

Bignell, G.R., Santarius, T., Pole, J.C.M., Butler, A.P., Perry, J., Pleasance, E., Greenman, C., Menzies, A., Taylor, S., Edkins, S.*, et al.* (2007). Architectures of somatic genomic rearrangement in human cancer amplicons at sequence-level resolution. Genome Research *17*, 1296-1303.

Birnbaum, D.J., Adélaïde, J., Mamessier, E., Finetti, P., Lagarde, A., Monges, G., Viret, F., Gonçalvès, A., Turrini, O., Delpero, J.-R.*, et al.* (2011). Genome profiling of pancreatic adenocarcinoma. Genes, Chromosomes and Cancer *50*, 456-465.

Bunting, S.F., and Nussenzweig, A. (2013). End-joining, translocations and cancer. Nature Reviews Cancer *13*, 443-454.

Caldas, C., Hahn, S.A., da Costa, L.T., Redston, M.S., Schutte, M., Seymour, A.B., Weinstein, C.L., Hruban, R.H., Yeo, C.J., and Kern, S.E. (1994). Frequent somatic mutations and homozygous deletions of the p16 (MTS1) gene in pancreatic adenocarcinoma. Nature Genetics *8*, 27-32.

Campbell, P.J., Stephens, P.J., Pleasance, E.D., O'Meara, S., Li, H., Santarius, T., Stebbings, L.A., Leroy, C., Edkins, S., Hardy, C.*, et al.* (2008). Identification of somatically acquired rearrangements in cancer using genome-wide massively parallel paired-end sequencing. Nature Genetics *40*, 722-729.

Campbell, P.J., Yachida, S., Mudie, L.J., Stephens, P.J., Pleasance, E.D., Stebbings, L.A., Morsberger, L.A., Latimer, C., McLaren, S., Lin, M.-L.*, et al.* (2010). The patterns and dynamics of genomic instability in metastatic pancreatic cancer. Nature *467*, 1109-1113.

Cappuzzo, F., Hirsch, F.R., Rossi, E., Bartolini, S., Ceresoli, G.L., Bemis, L., Haney, J., Witta, S., Danenberg, K., Domenichini, I.*, et al.* (2005). Epidermal growth factor receptor gene and protein and gefitinib sensitivity in non-small-cell lung cancer. Journal of the National Cancer Institute *97*, 643-655.

Chen, J.-M. (2001). Genomic rearrangements: Mutational mechanisms. In eLS (John Wiley & Sons, Ltd).

Chen, K., Wallis, J.W., McLellan, M.D., Larson, D.E., Kalicki, J.M., Pohl, C.S., McGrath, S.D., Wendl, M.C., Zhang, Q., Locke, D.P.*, et al.* (2009). BreakDancer: An algorithm for high-resolution mapping of genomic structural variation. Nature Methods *6*, 677-681.

Conrad, D.F., Bird, C., Blackburne, B., Lindsay, S., Mamanova, L., Lee, C., Turner, D.J., and Hurles, M.E. (2010). Mutation spectrum revealed by breakpoint sequencing of human germline CNVs. Nature Genetics *42*, 385-391.

Corbin, S., Neilly, M.E., Espinosa, R., 3rd, Davis, E.M., McKeithan, T.W., and Le Beau, M.M. (2002). Identification of unstable sequences within the common fragile site at 3p14.2: implications for the mechanism of deletions within fragile histidine triad gene/common fragile site at 3p14.2 in tumors. Cancer Research *62*, 3477-3484.

D'Andrea, A.D. (2003). The Fanconi Anemia/BRCA signaling pathway: disruption in cisplatin-sensitive ovarian cancers. Cell Cycle (Georgetown, Tex) *2*, 290-292.

Debniak, T., Gorski, B., Cybulski, C., Jakubowska, A., Kurzawski, G., Kladny, J., and Lubinski, J. (2001). Comparison of Alu-PCR, microsatelite instability, and immunohistochemical analyses in finding features characteristic for hereditary nonpolyposis colorectal cancer. Journal of Cancer Research and Clinical Oncology *127*, 565-569.

Depil, S., Roche, C., Dussart, P., and Prin, L. (2002). Expression of a human endogenous retrovirus, HERV-K, in the blood cells of leukemia patients. Leukemia *16*, 254-259.

Dhami, P., Coffey, A.J., Abbs, S., Vermeesch, J.R., Dumanski, J.P., Woodward, K.J., Andrews, R.M., Langford, C., and Vetrie, D. (2005). Exon array CGH: Detection of copy-number changes at the resolution of individual exons in the human genome. American Journal of Human Genetics *76*, 750-762.

Diehl, F., Li, M., Dressman, D., He, Y., Shen, D., Szabo, S., Diaz, L.A., Goodman, S.N., David, K.A., Juhl, H.*, et al.* (2005). Detection and quantification of mutations in the plasma of patients with colorectal tumors. Proceedings of the National Academy of Sciences of the United States of America *102*, 16368-16373.

Dillon, L.W., Burrow, A.A., and Wang, Y.H. (2010). DNA instability at chromosomal fragile sites in cancer. Current Genomics *11*, 326-337.

Drier, Y., Lawrence, M.S., Carter, S.L., Stewart, C., Gabriel, S.B., Lander, E.S., Meyerson, M., Beroukhim, R., and Getz, G. (2013). Somatic rearrangements across cancer reveal classes of samples with distinct patterns of DNA breakage and rearrangement-induced hypermutability. Genome Research *23*, 228-235.

Duffy, M.J. (1998). CA 19-9 as a marker for gastrointestinal cancers: a review. Annals Clinical Biochemistry *35 (Pt 3)*, 364-370.

Duffy, M.J., Sturgeon, C., Lamerz, R., Haglund, C., Holubec, V.L., Klapdor, R., Nicolini, A., Topolcan, O., and Heinemann, V. (2010). Tumor markers in pancreatic cancer: a European Group on Tumor Markers (EGTM) status report. Annals of Oncology *21*, 441-447.

Durkin, S.G., Ragland, R.L., Arlt, M.F., Mulle, J.G., Warren, S.T., and Glover, T.W. (2008). Replication stress induces tumor-like microdeletions in FHIT/FRA3B. Proceedings of the National Academy of Sciences of the United States of America *105*, 246-251.

Edkins, S., O'Meara, S., Parker, A., Stevens, C., Reis, M., Jones, S., Greenman, C., Davies, H., Dalgliesh, G., Forbes, S.*, et al.* (2006). Recurrent KRAS codon 146 mutations in human colorectal cancer. Cancer Biology & Therapy *5*, 928-932.

Farmer, H., McCabe, N., Lord, C.J., Tutt, A.N., Johnson, D.A., Richardson, T.B., Santarosa, M., Dillon, K.J., Hickson, I., Knights, C.*, et al.* (2005). Targeting the DNA repair defect in BRCA mutant cells as a therapeutic strategy. Nature *434*, 917-921.

Feuk, L., Carson, A.R., and Scherer, S.W. (2006). Structural variation in the human genome. Nature Reviews Genetics *7*, 85-97.

Finnegan, D.J. (1989). Eukaryotic transposable elements and genome evolution. Trends in genetics *5*, 103-107.

Forshew, T., Murtaza, M., Parkinson, C., Gale, D., Tsui, D.W.Y., Kaper, F., Dawson, S.J., Piskorz, A.M., Jimenez-Linan, M., Bentley, D.*, et al.* (2012). Noninvasive identification and monitoring of cancer mutations by targeted deep sequencing of plasma DNA. Science Translational Medicine *4*.

Fungtammasan, A., Walsh, E., Chiaromonte, F., Eckert, K.A., and Makova, K.D. (2012). A genome-wide analysis of common fragile sites: what features determine chromosomal instability in the human genome? Genome Research *22*, 993-1005.

Gandhi, M., Dillon, L.W., Pramanik, S., Nikiforov, Y.E., and Wang, Y.H. (2010). DNA breaks at fragile sites generate oncogenic RET/PTC rearrangements in human thyroid cells. Oncogene *29*, 2272-2280.

Glover, T.W., and Stein, C.K. (1988). Chromosome breakage and recombination at fragile sites. American Journal of Human Genetics *43*, 265-273.

Grabherr, M.G., Haas, B.J., Yassour, M., Levin, J.Z., Thompson, D.A., Amit, I., Adiconis, X., Fan, L., Raychowdhury, R., Zeng, Q.*, et al.* (2011). Full-length transcriptome assembly from RNA-Seq data without a reference genome. Nature Biotechnology *29*, 644-652.

Gudmundsdottir, K., and Ashworth, A. (2006). The roles of BRCA1 and BRCA2 and associated proteins in the maintenance of genomic stability. Oncogene *25*, 5864-5874.

Hahn, S.A., Schutte, M., Hoque, A., Moskaluk, C.A., daCosta, L.T., Rozenblum, E., Weinstein, C.L., Fischer, A., Yeo, C.J., Hruban, R.H.*, et al.* (1996). DPC4, a candidate tumor suppressor gene at human chromosome 18q21.1. Science *271*, 350-353.

Hastings, P.J., Lupski, J.R., Rosenberg, S.M., and Ira, G. (2009). Mechanisms of change in gene copy number. Nature Reviews Genetics *10*, 551-564.

Heim, S., and Mitelman, F. (1992). Cytogenetics of solid tumours. Recent Adv. Histopathology *15*, 37-66.

Heldin, C.H., Miyazono, K., and tenDijke, P. (1997). TGF-beta signalling from cell membrane to nucleus through SMAD proteins. Nature *390*, 465-471.

Helman, E., Lawrence, M.S., Stewart, C., Sougnez, C., Getz, G., and Meyerson, M. (2014). Somatic retrotransposition in human cancer revealed by whole-genome and exome sequencing. Genome Research *24*, 1053-1063.

Heyries, K.A., Tropini, C., Vaninsberghe, M., Doolin, C., Petriv, O.I., Singhal, A., Leung, K., Hughesman, C.B., and Hansen, C.L. (2011). Megapixel digital PCR. Nature Methods *8*, 649-651.

Hillmer, A.M., Yao, F., Inaki, K., Lee, W.H., Ariyaratne, P.N., Teo, A.S.M., Woo, X.Y., Zhang, Z., Zhao, H., Ukil, L.*, et al.* (2011). Comprehensive long-span paired-end-tag mapping reveals characteristic patterns of structural variations in epithelial cancer genomes. Genome Research.

Hormozdiari, F., Alkan, C., Eichler, E.E., and Sahinalp, S.C. (2009). Combinatorial algorithms for structural variation detection in high-throughput sequenced genomes. Genome Research *19*, 1270-1278.

Iafrate, A.J., Feuk, L., Rivera, M.N., Listewnik, M.L., Donahoe, P.K., Qi, Y., Scherer, S.W., and Lee, C. (2004). Detection of large-scale variation in the human genome. Nature Genetics *36*, 949-951.

Inoue, H., Ishii, H., Alder, H., Snyder, E., Druck, T., Huebner, K., and Croce, C.M. (1997). Sequence of the FRA3B common fragile region: implications for the mechanism of FHIT deletion. Proceedings of the National Academy of Sciences of the United States of America *94*, 14584-14589.

Ishioka, C., Suzuki, T., FitzGerald, M., Krainer, M., Shimodaira, H., Shimada, A., Nomizu, T., Isselbacher, K.J., Haber, D., and Kanamaru, R. (1997). Detection of heterozygous truncating mutations in the BRCA1 and APC genes by using a rapid screening assay in yeast. Proceedings of the National Academy of Sciences of the United States of America *94*, 2449-2453.

Jemal, A., Siegel, R., Xu, J., and Ward, E. (2010). Cancer Statistics, 2010. CA: A Cancer Journal for Clinicians *60*, 277-300.

Jiang, Y., Lucas, I., Young, D.J., Davis, E.M., Karrison, T., Rest, J.S., and Le Beau, M.M. (2009). Common fragile sites are characterized by histone hypoacetylation. Human Molecular Genetics *18*, 4501-4512.

Jiang, Y., Wang, Y., and Brudno, M. (2012). PRISM: pair-read informed split-read mapping for base-pair level detection of insertion, deletion and structural variants. Bioinformatics (Oxford, England) *28*, 2576-2583.

Jones, S., Zhang, X., Parsons, D.W., Lin, J.C.-H., Leary, R.J., Angenendt, P., Mankoo, P., Carter, H., Kamiyama, H., Jimeno, A.*, et al.* (2008). Core signaling pathways in human pancreatic cancers revealed by global genomic analyses. Science *321*, 1801-1806.

Kahlert, C., Melo, S.A., Protopopov, A., Tang, J., Seth, S., Koch, M., Zhang, J., Weitz, J., Chin, L., Futreal, A.*, et al.* (2014). Identification of double-stranded genomic DNA spanning all chromosomes with mutated KRAS and p53 DNA in the serum exosomes of patients with pancreatic cancer. The Journal of Biological Chemistry *289*, 3869-3875.

Kalthoff, H., Schmiegel, W., Roeder, C., Kasche, D., Schmidt, A., Lauer, G., Thiele, H.G., Honold, G., Pantel, K., Riethmuller, G.*, et al.* (1993). p53 and KRAS alterations in pancreatic epithelial-cell lesions Oncogene *8*, 289-298.

Kandoth, C., McLellan, M.D., Vandin, F., Ye, K., Niu, B., Lu, C., Xie, M., Zhang, Q., McMichael, J.F., Wyczalkowski, M.A.*, et al.* (2013). Mutational landscape and significance across 12 major cancer types. Nature *502*, 333-339.

Kassahn, K.S., Holmes, O., Nones, K., Patch, A.M., Miller, D.K., Christ, A.N., Harliwong, I., Bruxner, T.J., Xu, Q., Anderson, M.*, et al.* (2013). Somatic point mutation calling in low cellularity tumors. PLoS One *8*, e74380.

Kent, W.J. (2002). BLAT--the BLAST-like alignment tool. Genome Research *12*, 656-664.

Kidd, J.M., Cooper, G.M., Donahue, W.F., Hayden, H.S., Sampas, N., Graves, T., Hansen, N., Teague, B., Alkan, C., Antonacci, F.*, et al.* (2008). Mapping and sequencing of structural variation from eight human genomes. Nature *453*, 56-64.

Kidd, J.M., Graves, T., Newman, T.L., Fulton, R., Hayden, H.S., Malig, M., Kallicki, J., Kaul, R., Wilson, R.K., and Eichler, E.E. (2010). A human genome structural variation sequencing resource reveals insights into mutational mechanisms. Cell *143*, 837-847.

Klapman, J., and Malafa, M.P. (2008). Early detection of pancreatic cancer: why, who, and how to screen. Cancer Control *15*, 280-287.

Kloosterman, W.P., Hoogstraat, M., Paling, O., Tavakoli-Yaraki, M., Renkens, I., Vermaat, J.S., van Roosmalen, M.J., van Lieshout, S., Nijman, I.J., Roessingh, W.*, et al.* (2011). Chromothripsis is a common mechanism driving genomic rearrangements in primary and metastatic colorectal cancer. Genome Biology *12*, R103.

Knudson, A.G. (1971). Mutation and cancer: Statistical study of retinoblastoma. Proceedings of the National Academy of Sciences of the United States of America *68*, 820-823.

Koprowski, H., Steplewski, Z., Mitchell, K., Herlyn, M., Herlyn, D., and Fuhrer, P. (1979). Colorectal carcinoma antigens detected by hybridoma antibodies. Somatic Cell Genetics *5*, 957-971.

Korbel, J.O., Abyzov, A., Mu, X.J., Carriero, N., Cayting, P., Zhang, Z., Snyder, M., and Gerstein, M.B. (2009). PEMer: a computational framework with simulation-based error models for inferring genomic structural variants from massive paired-end sequencing data. Genome Biology *10*, R23.

Korbel, J.O., and Campbell, P.J. (2013). Criteria for inference of chromothripsis in cancer genomes. Cell *152*, 1226-1236.

Korbel, J.O., Urban, A.E., Affourtit, J.P., Godwin, B., Grubert, F., Simons, J.F., Kim, P.M., Palejev, D., Carriero, N.J., Du, L.*, et al.* (2007). Paired-end mapping reveals extensive structural variation in the human genome. Science *318*, 420-426.

Krzywinski, M., Schein, J., Birol, I., Connors, J., Gascoyne, R., Horsman, D., Jones, S.J., and Marra, M.A. (2009). Circos: an information aesthetic for comparative genomics. Genome Research *19*, 1639-1645.

Lam, H.Y., Mu, X.J., Stutz, A.M., Tanzer, A., Cayting, P.D., Snyder, M., Kim, P.M., Korbel, J.O., and Gerstein, M.B. (2010). Nucleotide resolution analysis of structural variants using BreakSeq and a breakpoint library. Nature Biotechnology *28*, 47-55.

Leary, R.J., Kinde, I., Diehl, F., Schmidt, K., Clouser, C., Duncan, C., Antipova, A., Lee, C., McKernan, K., De La Vega, F.M.*, et al.* (2010). Development of personalized tumor biomarkers using massively parallel sequencing. Science Translational Medicine *2*, 20ra14.

Leary, R.J., Lin, J.C., Cummins, J., Boca, S., Wood, L.D., Parsons, D.W., Jones, S.n., Sj√∂blom, T., Park, B.-H., Parsons, R.*, et al.* (2008). Integrated analysis of homozygous deletions, focal amplifications, and sequence alterations in breast and colorectal cancers.

Proceedings of the National Academy of Sciences of the United States of America *105*, 16224-16229.

Leary, R.J., Sausen, M., Kinde, I., Papadopoulos, N., Carpten, J.D., Craig, D., O'Shaughnessy, J., Kinzler, K.W., Parmigiani, G., Vogelstein, B.*, et al.* (2012). Detection of chromosomal alterations in the circulation of cancer patients with whole-genome sequencing. Science Translation Medicine *4*, 162-154.

Lee, E., Iskow, R., Yang, L., Gokcumen, O., Haseley, P., Luquette, L.J., 3rd, Lohr, J.G., Harris, C.C., Ding, L., Wilson, R.K.*, et al.* (2012). Landscape of somatic retrotransposition in human cancers. Science *337*, 967-971.

Lee, J.A., Carvalho, C.M., and Lupski, J.R. (2007). A DNA replication mechanism for generating nonrecurrent rearrangements associated with genomic disorders. Cell *131*, 1235-1247.

Levine, A.J. (1997). p53, the cellular gatekeeper for growth and division. Cell *88*, 323-331.

Levy, S., Sutton, G., Ng, P.C., Feuk, L., Halpern, A.L., Walenz, B.P., Axelrod, N., Huang, J., Kirkness, E.F., Denisov, G.*, et al.* (2007). The diploid genome sequence of an individual human. PLoS Biology *5*, e254.

Li, H., and Durbin, R. (2009). Fast and accurate short read alignment with Burrows-Wheeler transform. Bioinformatics (Oxford, England) *25*, 1754-1760.

Li, H., Handsaker, B., Wysoker, A., Fennell, T., Ruan, J., Homer, N., Marth, G., Abecasis, G., and Durbin, R. (2009). The sequence alignment/map format and SAMtools. Bioinformatics (Oxford, England) *25*, 2078-2079.

Liu, G., Yang, D., Sun, Y., Shmulevich, I., Xue, F., Sood, A.K., and Zhang, W. (2012). Differing clinical impact of BRCA1 and BRCA2 mutations in serous ovarian cancer. Pharmacogenomics *13*, 1523-1535.

Liu, Y., and West, S.C. (2002). Distinct functions of BRCA1 and BRCA2 in double-strand break repair. Breast Cancer Research *4*, 9-13.

Lo, A.W., Sabatier, L., Fouladi, B., Pottier, G., Ricoul, M., and Murnane, J.P. (2002). DNA amplification by breakage/fusion/bridge cycles initiated by spontaneous telomere loss in a human cancer cell line. Neoplasia (New York, NY) *4*, 531-538.

Makela, T.P., Saksela, K., and Alitalo, K. (1992). Amplification and rearrangement of L-myc in human small-cell lung cancer. Mutation Research *276*, 307-315.

Mauillon, J.L., Michel, P., Limacher, J.M., Latouche, J.B., Dechelotte, P., Charbonnier, F., Martin, C., Moreau, V., Metayer, J., Paillot, B.*, et al.* (1996). Identification of novel germline hMLH1 mutations including a 22 kb Alu-mediated deletion in patients with familial colorectal cancer. Cancer Research *56*, 5728-5733.

McBride, D.J., Orpana, A.K., Sotiriou, C., Joensuu, H., Stephens, P.J., Mudie, L.J., Hamalainen, E., Stebbings, L.A., Andersson, L.C., Flanagan, A.M.*, et al.* (2010). Use of cancer-specific genomic rearrangements to quantify disease burden in plasma from patients with solid tumors. Genes, Chromosomes & Cancer *49*, 1062-1069.

McKernan, K.J., Peckham, H.E., Costa, G.L., McLaughlin, S.F., Fu, Y., Tsung, E.F., Clouser, C.R., Duncan, C., Ichikawa, J.K., Lee, C.C.*, et al.* (2009). Sequence and structural variation in a human genome uncovered by short-read, massively parallel ligation sequencing using two-base encoding. Genome Research *19*, 1527-1541.

McNeil, N., and Ried, T. (2000). Novel molecular cytogenetic techniques for identifying complex chromosomal rearrangements: technology and applications in molecular medicine. Expert Reviews in Molecular Medicine *2000*, 1-14.

Medvedev, P., Stanciu, M., and Brudno, M. (2009). Computational methods for discovering structural variation with next-generation sequencing. Nature Methods *6*, S13-S20.

Milbury, C.A., Correll, M., Quackenbush, J., Rubio, R., and Makrigiorgos, G.M. (2012). COLD-PCR enrichment of rare cancer mutations prior to targeted amplicon resequencing. Clinical Chemistry *58*, 580-589.

Montagna, M., Santacatterina, M., Torri, A., Menin, C., Zullato, D., Chieco-Bianchi, L., and D'Andrea, E. (1999). Identification of a 3 kb Alu-mediated BRCA1 gene rearrangement in two breast/ovarian cancer families. Oncogene *18*, 4160-4165.

Murtaza, M., Dawson, S.J., Tsui, D.W., Gale, D., Forshew, T., Piskorz, A.M., Parkinson, C., Chin, S.F., Kingsbury, Z., Wong, A.S.*, et al.* (2013). Non-invasive analysis of acquired resistance to cancer therapy by sequencing of plasma DNA. Nature *497*, 108-112.

Narayan, A., Carriero, N.J., Gettinger, S.N., Kluytenaar, J., Kozak, K.R., Yock, T.I., Muscato, N.E., Ugarelli, P., Decker, R.H., and Patel, A.A. (2012). Ultrasensitive measurement of hotspot mutations in tumor DNA in blood using error-suppressed multiplexed deep sequencing. Cancer Research *72*, 3492-3498.

Negrini, S., Gorgoulis, V.G., and Halazonetis, T.D. (2010). Genomic instability - An evolving hallmark of cancer. Nature Reviews Molecular Cell Biology *11*, 220-228.

Neill, N., Torchia, B., Bejjani, B., Shaffer, L., and Ballif, B. (2010). Comparative analysis of copy number detection by whole-genome BAC and oligonucleotide array CGH. Molecular Cytogenetics *3*, 11.

Newman, A.M., Bratman, S.V., To, J., Wynne, J.F., Eclov, N.C., Modlin, L.A., Liu, C.L., Neal, J.W., Wakelee, H.A., Merritt, R.E.*, et al.* (2014). An ultrasensitive method for quantitating circulating tumor DNA with broad patient coverage. Nature Medicine *20*, 548-554.

Ng, C.K., Cooke, S.L., Howe, K., Newman, S., Xian, J., Temple, J., Batty, E.M., Pole, J.C., Langdon, S.P., Edwards, P.A.*, et al.* (2012). The role of tandem duplicator phenotype in tumour evolution in high-grade serous ovarian cancer. Journal of Pathology *226*, 703-712.

Nik-Zainal, S., Alexandrov, L.B., Wedge, D.C., Van Loo, P., Greenman, C.D., Raine, K., Jones, D., Hinton, J., Marshall, J., Stebbings, L.A.*, et al.* (2012). Mutational processes molding the genomes of 21 breast cancers. Cell *149*, 979-993.

Nowell, P.C., and Hungerford, D.A. (1960). Minute chromosome in human chronic granulocytic leukemia. Science *132*, 1497-1497.

Nussey, S., and Whitehead, S. (2001). In Endocrinology: An integrated approach (Oxford: BIOS Scientific Publishers Limited).

Onishi-Seebacher, M., and Korbel, J.O. (2011). Challenges in studying genomic structural variant formation mechanisms: the short-read dilemma and beyond. BioEssays *33*, 840-850.

Patel, A.G., Sarkaria, J.N., and Kaufmann, S.H. (2011). Nonhomologous end joining drives poly(ADP-ribose) polymerase (PARP) inhibitor lethality in homologous recombination-deficient cells. Proceedings of the National Academy of Sciences of the United States of America *108*, 3406-3411.

Pellegata, N.S., Sessa, F., Renault, B., Bonato, M., Leone, B.E., Solcia, E., and Ranzani, G.N. (1994). K-ras and p53 gene mutations in pancreatic cancer: Ductal and nonductal tumors progress through different genetic lesions. Cancer Research *54*, 1556-1560.

Pfeiffer, P., Goedecke, W., and Obe, G. (2000). Mechanisms of DNA double-strand break repair and their potential to induce chromosomal aberrations. Mutagenesis *15*, 289-302.

Pinheiro, L.B., Coleman, V.A., Hindson, C.M., Herrmann, J., Hindson, B.J., Bhat, S., and Emslie, K.R. (2012). Evaluation of a droplet digital polymerase chain reaction format for DNA copy number quantification. Analytical chemistry *84*, 1003-1011.

Pinkel, D., Segraves, R., Sudar, D., Clark, S., Poole, I., Kowbel, D., Collins, C., Kuo, W.-L., Chen, C., Zhai, Y.*, et al.* (1998). High resolution analysis of DNA copy number variation using comparative genomic hybridization to microarrays. Nature Genetics *20*, 207-211.

Popescu, N.C. (2003). Genetic alterations in cancer as a result of breakage at fragile sites. Cancer Letters *192*, 1-17.

Popova, T., Manie, E., Rieunier, G., Caux-Moncoutier, V., Tirapo, C., Dubois, T., Delattre, O., Sigal-Zafrani, B., Bollet, M., Longy, M.*, et al.* (2012). Ploidy and large-scale genomic instability consistently identify basal-like breast carcinomas with BRCA1/2 inactivation. Cancer Research *72*, 5454-5462.

Popova, T., Manie, E., Stoppa-Lyonnet, D., Rigaill, G., Barillot, E., and Stern, M.H. (2009). Genome Alteration Print (GAP): a tool to visualize and mine complex cancer genomic profiles obtained by SNP arrays. Genome Biology *10*, R128.

Prak, E.T., and Kazazian, H.H., Jr. (2000). Mobile elements and the human genome. Nature Reviews Genetics *1*, 134-144.

Quek, K., Nones, K., Patch, A.M., Fink, J.L., Newell, F., Cloonan, N., Miller, D., Fadlullah, M.Z., Kassahn, K., Christ, A.N.*, et al.* (2014). A workflow to increase verification rate of chromosomal structural rearrangements using high-throughput next-generation sequencing. BioTechniques *57*, 31-38.

Raimondi, S., Maisonneuve, P., and Lowenfels, A.B. (2009). Epidemiology of pancreatic cancer: an overview. Nature Reviews Gastroenterology & Hepatology *6*, 699-708.

Rausch, T., Zichner, T., Schlattl, A., Stutz, A.M., Benes, V., and Korbel, J.O. (2012). DELLY: structural variant discovery by integrated paired-end and split-read analysis. Bioinformatics (Oxford, England) *28*, i333-i339.

Ribeiro, G.R., Francisco, G., Teixeira, L.V., Romao-Correia, R.F., Sanches, J.A., Jr., Neto, C.F., and Ruiz, I.R. (2004). Repetitive DNA alterations in human skin cancers. Journal of Dermatological Science *36*, 79-86.

Rowe, S.M., Coughlan, S.J., McKenna, N.J., Garrett, E., Kieback, D.G., Carney, D.N., and Headon, D.R. (1995). Ovarian carcinoma-associated TaqI restriction fragment length polymorphism in intron G of the progesterone receptor gene is due to an Alu sequence insertion. Cancer Research *55*, 2743-2745.

Rusk, N., and Kiermer, V. (2008). Primer: Sequencing--the next generation. Nature Methods *5*, 15.

Sadikovic, B., Al-Romaih, K., Squire, J.A., and Zielenska, M. (2008). Cause and consequences of genetic and epigenetic alterations in human cancer. Current Genomics *9*, 394-408.

Saif, M.W. (2006). Pancreatic cancer: highlights from the 42nd annual meeting of the American Society of Clinical Oncology, 2006. JOP *7*, 337-348.

Schroder, J., Hsu, A., Boyle, S.E., Macintyre, G., Cmero, M., Tothill, R.W., Johnstone, R.W., Shackleton, M., and Papenfuss, A.T. (2014). Socrates: identification of genomic rearrangements in tumour genomes by re-aligning soft clipped reads. Bioinformatics (Oxford, England).

Schutte, M., Hruban, R.H., Geradts, J., Maynard, R., Hilgers, W., Rabindran, S.K., Moskaluk, C.A., Hahn, S.A., SchwarteWaldhoff, I., Schmiegel, W.*, et al.* (1997). Abrogation of the Rb/p16 tumor-suppressive pathway in virtually all pancreatic carcinomas. Cancer Research *57*, 3126-3130.

Schwarzenbach, H., Hoon, D.S., and Pantel, K. (2011). Cell-free nucleic acids as biomarkers in cancer patients. Nature Reviews Cancer *11*, 426-437.

Shann, Y.J., Cheng, C., Chiao, C.H., Chen, D.T., Li, P.H., and Hsu, M.T. (2008). Genome-wide mapping and characterization of hypomethylated sites in human tissues and breast cancer cell lines. Genome Research *18*, 791-801.

Shendure, J., and Ji, H. (2008). Next-generation DNA sequencing. Nature Biotechnology *26*, 1135-1145.

Sherr, C.J. (1996). Cancer Cell Cycles. Science *274*, 1672-1677.

Slotkin, R.K., and Martienssen, R. (2007). Transposable elements and the epigenetic regulation of the genome. Nature Reviews Genetics *8*, 272-285.

Soda, M., Choi, Y.L., Enomoto, M., Takada, S., Yamashita, Y., Ishikawa, S., Fujiwara, S.-i., Watanabe, H., Kurashina, K., Hatanaka, H.*, et al.* (2007). Identification of the transforming EML4-ALK fusion gene in non-small-cell lung cancer. Nature *448*, 561-566.

Solyom, S., Ewing, A.D., Rahrmann, E.P., Doucet, T., Nelson, H.H., Burns, M.B., Harris, R.S., Sigmon, D.F., Casella, A., Erlanger, B.*, et al.* (2012). Extensive somatic L1 retrotransposition in colorectal tumors. Genome Research *22*, 2328-2338.

Song, S., Nones, K., Miller, D., Harliwong, I., Kassahn, K.S., Pinese, M., Pajic, M., Gill, A.J., Johns, A.L., Anderson, M.*, et al.* (2012). qpure: A tool to estimate tumor cellularity from genome-wide single-nucleotide polymorphism profiles. PLoS One *7*, e45835.

Sorenson, G.D., Pribish, D.M., Valone, F.H., Memoli, V.A., Bzik, D.J., and Yao, S.L. (1994). Soluble normal and mutated DNA sequences from single copy genes in human blood. Cancer Epidemiology Biomarkers & Prevention *3*, 67-71.

Steinberg, W. (1990). The clinical utility of the CA 19-9 tumor-associated antigen. American Journal of Gastroenterology *85*, 350-355.

Stephens, P.J., Greenman, C.D., Fu, B.Y., Yang, F.T., Bignell, G.R., Mudie, L.J., Pleasance, E.D., Lau, K.W., Beare, D., Stebbings, L.A.*, et al.* (2011). Massive genomic rearrangement acquired in a single catastrophic event during cancer development. Cell *144*, 27-40.

Stephens, P.J., McBride, D.J., Lin, M.-L., Varela, I., Pleasance, E.D., Simpson, J.T., Stebbings, L.A., Leroy, C., Edkins, S., Mudie, L.J.*, et al.* (2009). Complex landscapes of somatic rearrangement in human breast cancer genomes. Nature *462*, 1005-1010.

Stratton, M.R., Campbell, P.J., and Futreal, P.A. (2009). The cancer genome. Nature *458*, 719-724.

Strout, M.P., Marcucci, G., Bloomfield, C.D., and Caligiuri, M.A. (1998). The partial tandem duplication of ALL1 (MLL) is consistently generated by Alu-mediated homologous recombination in acute myeloid leukemia. Proceedings of the National Academy of Sciences of the United States of America *95*, 2390-2395.

Sykes, P.J., Neoh, S.H., Brisco, M.J., Hughes, E., Condon, J., and Morley, A.A. (1992). Quantitation of targets for PCR by use of limiting dilution. BioTechniques *13*, 444-449.

Tagawa, H., Karnan, S., Suzuki, R., Matsuo, K., Zhang, X., Ota, A., Morishima, Y., Nakamura, S., and Seto, M. (2004). Genome-wide array-based CGH for mantle cell lymphoma: identification of homozygous deletions of the proapoptotic gene BIM. Oncogene *24*, 1348-1358.

Takasaki, H., Uchida, E., Tempero, M.A., Burnett, D.A., Metzgar, R.S., and Pour, P.M. (1988). Correlative study on expression of CA 19-9 and DU-PAN-2 in tumor tissue and in serum of pancreatic cancer patients. Cancer Research *48*, 1435-1438.

Taylor, D.D., and Gercel-Taylor, C. (2008). MicroRNA signatures of tumor-derived exosomes as diagnostic biomarkers of ovarian cancer. Gynecologic Oncology *110*, 13-21.

Thomas, R.K., Baker, A.C., Debiasi, R.M., Winckler, W., Laframboise, T., Lin, W.M., Wang, M., Feng, W., Zander, T., MacConaill, L.*, et al.* (2007). High-throughput oncogene mutation profiling in human cancer. Nature Genetics *39*, 347-351.

Totoki, Y., Tatsuno, K., Yamamoto, S., Arai, Y., Hosoda, F., Ishikawa, S., Tsutsumi, S., Sonoda, K., Totsuka, H., Shirakihara, T.*, et al.* (2011). High-resolution characterization of a hepatocellular carcinoma genome. Nature Genetics *43*, 464-469.

Tuzun, E., Sharp, A.J., Bailey, J.A., Kaul, R., Morrison, V.A., Pertz, L.M., Haugen, E., Hayden, H., Albertson, D., Pinkel, D.*, et al.* (2005). Fine-scale structural variation of the human genome. Nature Genetics *37*, 727-732.

Vasioukhin, V., Anker, P., Maurice, P., Lyautey, J., Lederrey, C., and Stroun, M. (1994). Point mutations of the N-RAS gene in the blood plasma DNA of patients with myelodysplastic syndrome or acute myelogenous leukemia British Journal of Haematology *86*, 774-779.

Villarreal, D.D., Lee, K., Deem, A., Shim, E.Y., Malkova, A., and Lee, S.E. (2012). Microhomology directs diverse DNA break repair pathways and chromosomal translocations. PLoS Genetics *8*, e1003026.

Vogelstein, B., and Kinzler, K.W. (1999). Digital PCR. Proceedings of the National Academy of Sciences of the United States of America *96*, 9236-9241.

Waddell, N., Pajic, M., Patch, A.M., Chang, D.K., Kassahn, K.S., Bailey, P., Johns, A.L., Miller, D., Nones, K., Quek, K.*, et al.* (2015). Whole genomes redefine the mutational landscape of pancreatic cancer. Nature *518*, 495-501.

Wang, J., Mullighan, C.G., Easton, J., Roberts, S., Heatley, S.L., Ma, J., Rusch, M.C., Chen, K., Harris, C.C., Ding, L.*, et al.* (2011). CREST maps somatic structural variation in cancer genomes with base-pair resolution. Nature Methods *8*, 652-654.

Wang, L., Tsutsumi, S., Kawaguchi, T., Nagasaki, K., Tatsuno, K., Yamamoto, S., Sang, F., Sonoda, K., Sugawara, M., Saiura, A.*, et al.* (2012). Whole-exome sequencing of human pancreatic cancers and characterization of genomic instability caused by MLH1 haploinsufficiency and complete deficiency. Genome Research *22*, 208-219.

Warburton, D. (1991). De novo balanced chromosome rearrangements and extra marker chromosomes identified at prenatal diagnosis: clinical significance and distribution of breakpoints. American Journal of Human Genetics *49*, 995-1013.

Witcher, M., and Emerson, B.M. (2009). Epigenetic silencing of the p16(INK4a) tumor suppressor is associated with loss of CTCF binding and a chromatin boundary. Molecular Cell *34*, 271-284.

Xing, E.P., Yang, G.Y., Wang, L.D., Shi, S.T., and Yang, C.S. (1999). Loss of heterozygosity of the Rb gene correlates with pRb protein expression and associates with p53 alteration in human esophageal cancer. Clinical Cancer Research *5*, 1231-1240.

Xing, J., Zhang, Y., Han, K., Salem, A.H., Sen, S.K., Huff, C.D., Zhou, Q., Kirkness, E.F., Levy, S., Batzer, M.A.*, et al.* (2009). Mobile elements create structural variation: analysis of a complete human genome. Genome Research *19*, 1516-1526.

Yang, L., Luquette, L.J., Gehlenborg, N., Xi, R., Haseley, P.S., Hsieh, C.H., Zhang, C., Ren, X., Protopopov, A., Chin, L.*, et al.* (2013). Diverse mechanisms of somatic structural variations in human cancer genomes. Cell *153*, 919-929.

Yang, N., and Kazazian, H.H., Jr. (2006). L1 retrotransposition is suppressed by endogenously encoded small interfering RNAs in human cultured cells. Nature structural & Molecular Biology *13*, 763-771.

Yu, M., and Tannock, I.F. (2012). Targeting tumor architecture to favor drug penetration: a new weapon to combat chemoresistance in pancreatic cancer? Cancer cell *21*, 327-329.

Yung, T.K., Chan, K.C., Mok, T.S., Tong, J., To, K.F., and Lo, Y.M. (2009). Single-molecule detection of epidermal growth factor receptor mutations in plasma by microfluidics digital PCR in non-small cell lung cancer patients. Clinical Cancer Research *15*, 2076-2084.

Zhang, J., and Powell, S.N. (2005). The role of the BRCA1 tumor suppressor in DNA double-strand break repair. Molecular Cancer Research : MCR *3*, 531-539.